IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received December 30, 2020, accepted January 20, 2021, date of publication February 8, 2021, date of current version February 26, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3057900

Minimizing Excess Timing Guard Banding Under Transistor Self-Heating Through Biasing at Zero-Temperature Coefficient

SAMI SALAMIN[1], (Student Member, IEEE), VICTOR M. VAN SANTEN[2], (Member, IEEE), MARTIN RAPP[1], (Graduate Student Member, IEEE), JÖRG HENKEL[1], (Fellow, IEEE), AND HUSSAM AMROUCH[2], (Member, IEEE)

[1]Department of Computer Science, Karlsruhe Institute of Technology (KIT), 76131 Karlsruhe, Germany
[2]Department of Computer Science, University of Stuttgart, 70569 Stuttgart, Germany

Corresponding author: Sami Salamin (sami.salamin@kit.edu)

The work of Victor M. Van Santen and Hussam Amrouch was done in part at KIT.

ABSTRACT Self-Heating Effects (SHE) is known as one of the key reliability challenges in FinFET and beyond. Large timing guard bands are necessary, which we try to reduce. In this work, we propose operating (biasing) processors at Zero-Temperature Coefficient (ZTC) to contain (mitigate) SHE-induced delay. Operating at ZTC allows near-zero timing guard band to protect circuits against SHE. However, a trade-off is found between thermal timing guard band and performance loss from lowering the voltage.

INDEX TERMS Inverse-temperature dependence, positive-temperature dependence, self-heating effects, zero-temperature coefficient, reliability, guard band, timing.

I. INTRODUCTION

Fin Field-Effect Transistor (FinFET) devices are widely used, due to their reduced leakage and excellent subthreshold slope compared to planar MOSFET. FinFET advantages resulted from the new 3D structure of transistors with a vertical junction. The introduction of the FinFET 3D structure and due to the low thermal conductivity of the gate dielectric, the heat dissipation from a FinFET channel is limited overtime compared to planar MOSFETs as shown in Fig. 1. Moreover, since the thermal resistance (Rth) of the gate is high, the heat transport towards the body is limited. Hence, *most of the heat* generated within the FinFET transistor's channel remains within its channel as it slowly escapes to the body.

Self Heating Effect (SHE) refers to elevated channel temperatures (T_C) and their impact on the performance of the transistor. The channel temperature is elevated due to Joule heating by the current flow through the channel.

When SHE-induced T_C of the transistors in the circuit raises, I_D in the ON-state drops and hence increases delay of the transistor at nominal voltage, reducing the maximum clock frequency and thus circuit performance. At the same time, the leakage current (I_D in OFF-state) I_{off} increases

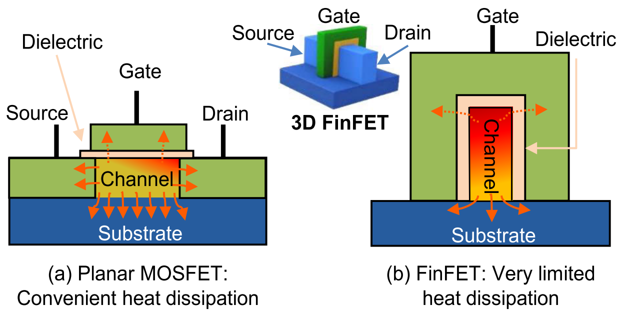

FIGURE 1. (a) Planar MOSFET transistor: Heat dissipation from the channel is convenient due to conduction towards the substrate. This allows MOSFET to dissipate most of the generated heat within its channel. (b) 3D FinFET (side view of the channel directly after drain to show the hotspot within channel): Exhibits limited heat dissipation from its channel to the body.

(due to strong impact of lower V_{th} due to temperature), thus increasing leakage power of the circuit [1].

Following the dependence between the operating voltage and temperature, three key regions exist: Positive-Temperature Dependence (PTD) (i.e., increasing T_C reduces I_D), Zero-Temperature Coefficient (ZTC) (i.e., increasing T_C does not change I_D) and Inverse-Temperature Dependence (ITD) (i.e., increasing T_C increases I_D) [2], [3] (more details Appendix A-C).

The associate editor coordinating the review of this manuscript and approving it for publication was Nagarajan Raghavan.

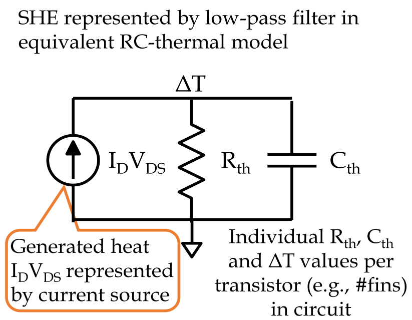

FIGURE 2. Schematic diagram of a SHE model represented by a low-pass filter in an equivalent RC-thermal model.

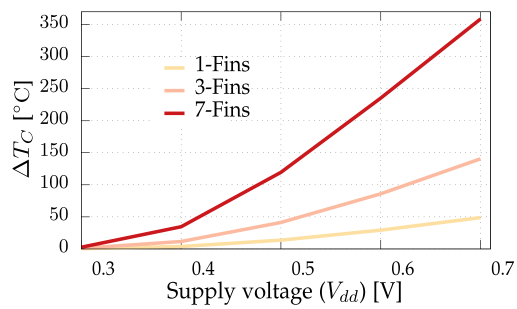

FIGURE 3. Temperature increase within the transistor's channel (ΔT_C) due to SHE over supply voltages V_{dd}. The results are generated by employing the SHE model in Fig. 2 for 1,3 and 7-Fins transistors at typical operating temperature (i.e., at room temperature of 25°C).

SHE is a fundamental result of the new transistor design (i.e., 3D structure) and we can only try to reduce its impacts to recover the lost performance. Therefore, we must reduce impact of the high SHE-induced T_C on the circuit. The ZTC operating point is well-suited to minimize SHE impacts on the circuit's delay. By definition it is a point (or region) where the temperature has little impact on the circuit's delay. Consequently, we propose to minimize the impact of SHE by operating at or near ZTC at the cost of operating at a lower V_{dd}. We shift V_{dd} to a voltage lower than nominal, which comes with its own *performance loss*.

Our novel contributions within this paper are:

(1) We are the first to analyze the impacts of SHE on both the timing and power of large digital circuits i.e., including a full microprocessor. For this purpose, we extend the existing Multi-Corner Multi-Mode (MCMM) approach used in EDA tool flow with SHE-aware cell libraries to enable SHE modeling for the entire chip.

(2) We show that the ZTC point depends on the topology of the circuit and thus differs for each circuit.

(3) For the first time, we operate circuits at near zero-temperature coefficient (N-ZTC). N-ZTC models the total temperature dependence of the circuit, consisting of more than just the average of the distinct ZTCs of the sub-circuits within the circuit. Operating at N-ZTC minimizes SHE-induced variance in performance and power.

(4) We qualitatively and quantitatively compare traditional timing guard banding with N-ZTC in terms of performance and energy of multi-core systems.

II. RELATED WORK

A large body of works studied simple circuits to characterize ITD, from single transistors to small circuits. The work in [4] studies the operation of transistors in different thermal regions. The work in [5] presents an analysis of ZTC of a 32-bit CMOS adder based on SPICE simulations at 65nm.

Reference [6] shows ITD impact on performance in a 65nm CMOS ring oscillator simulations using SPICE in the sub-threshold regime. This quantitative study shows that ZTC occurs at $V_{ZTC} = 0.9$V. However, studying ITD and ZTC in a single transistor or simple circuits is insufficient, because their ZTC is different and thus a single RO is not representative for a chip. For circuits, Intel presented in [7], a 130nm test chip containing different types of ring oscillators and found distinct V_{ZTC} in the range between 0.783-0.866V.

SHE is well studied at the transistor level since it is well known for Silicon-On-Insulator (SOI) devices [8] and power MOSFETs [9]. Recently, transistor-level studies in FinFETs provide a good understanding of SHE in transistors [10], [11] [12]. However, these studies are limited to simple circuits and the impact of SHE beyond ring oscillators and SRAM cells is not yet studied. Importantly, SHE can aggravate more reliability issues [13].

III. SELF HEATING MODELING

To study the impact of SHE on large circuits, we enhance and employ the standard EDA tools. Since SHE originally is analyzed at the transistor level, we start our analysis there. We perform single transistor SPICE simulations to determine $\Delta T_C(SHE)$ under different conditions (e.g., different V_{dd}, switching frequencies, number of fins, etc.).

Modeling Self-Heating Effects: In this work, we employ the model typically used in SPICE circuit simulations. It relies on a RC-thermal network to model SHE. The industry-standard FinFET compact model BSIM-CMG [14] uses this model to model SHE. With this model, $\Delta T_C(SHE)$ can be estimated by solving for the voltage at node T (T_C). Please note, BSIM-CMG model does not precisely capture all SHE impacts, which might slightly alter the delay results. The temporal behavior of SHE is given by the time constant $\tau_{th} = C_{th}{\cdot}R_{th}$. A large time constant (e.g., $\tau_{th} = 100$ns) result in slow heating/cooling of the channel, while fast time constants (e.g., $\tau_{th} = 0.5$ns) result in rapid temperature changes. Currently, typical time constants are approximately 1ns [15].

Transistor SHE Simulations: To model the electrical characteristics of pFinFET and nFinFET transistors, we employ the modelcard from the ASAP7 PDK [16]. The employed transistor model is BSIM-CMGv110 [14]. We perform simulations for pFinFET and nFinFET under a range of voltages and for different numbers of fins. We calibrated BSIM-CMG with 7nm FinFET SHE parameters from [11]. This could result in SHE underestimation due to lower I_d in ASAP7 [16] compared to [11] which we used to calibrate R_{th} and C_{th}. However, as the resulted ΔT_C is already high, we did not configure the transistor to have the same I_d as in [11] to stay optimistic. The simulation of a single transistor using typical operation conditions (i.e., 25°C $V_{dd} = 0.7$V) and 3 fins shows $\Delta T_C(SHE) \approx 150$°C. However, $\Delta T_C(SHE)$ significantly increased when we change number of fins to 7. Multiple fins heat the substrate and thus each other. Consequently, increasing the number of fins results in high temperatures (350°C shown in Fig. 3). Such a high T_C occurs under worst-case corner (continuous heating due to DC currents, high fin counts, high voltage). Note that worst case means the slowest delay always. Fig. 3 shows that $\Delta T_C(SHE)$ decreases with V_{dd} decreases and reaches ≈50°C at 0.5V for 3 fins and ≈120°C for 7 fins.

IV. MINIMIZING THERMAL DEPENDENCE VIA ZTC OPERATION IN LARGE CIRCUITS

We show here the key challenge behind finding single ZTC for large circuits, exceeding 100K transistors. Then we illustrate our approach in finding the point near ZTC with *minuscule* temperature-induced variance.

A. FINDING THE ZTC OF STANDARD CELLS

The V_{ZTC} is the supply voltage (V_{dd}) where ZTC is observed. Obtaining ZTC voltages for large circuits, such as a processor, while considering SHE is challenging. A microprocessor features thousands of subcircuits. Each contains many connected standard cells with a unique V_{ZTC} per cell type [7]. This is due to the different transistor types (e.g., more pFinFET than nFinFET in a particular cell) where each transistor type has a unique V_{ZTC} [4], different topology (transistors in series, transistors in parallel, etc.), and ultimately different transistor configurations (number of fins) per cell. Moreover, considering the different operating conditions of each cell creates a non-negligible variance in V_{ZTC}. To take the impact of the operating conditions into account, we consider 7 input signal slews (t_{slew}) along with 7 output load capacitances (C_{load}). These are typical values for industrial and academic cell library characterization [17]. Consequently, cell topology, t_{slew} and C_{load} result in various V_{ZTC} for different standard cells. The 7 × 7 propagation delay matrix for each standard cell is arranged as follow:

$$7 \times 7 = \begin{bmatrix} (t_{slew_1}, C_{load_1}) & \dots & (t_{slew_1}, C_{load_7}) \\ \vdots & \ddots & \vdots \\ (t_{slew_7}, C_{load_1}) & \dots & (t_{slew_7}, C_{load_7}) \end{bmatrix}$$

For example, to illustrate the variations in V_{ZTC} under SHE, the 7 × 7 of V_{ZTC} matrix of NANDx2 (nand gate) cell experiments for the average rise delay shows various V_{ZTC} as follow:

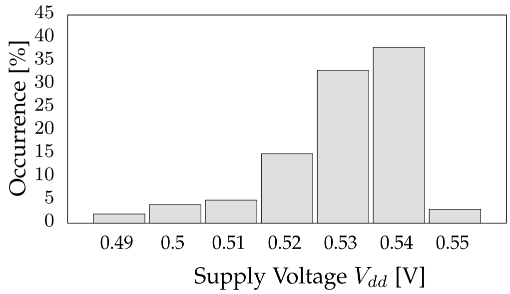

FIGURE 4. Histogram of the results of V_{ZTC} of cells. Experiments cover all operating conditions of all cells extracted by simulating every standard cell at high T_C (with SHE) and low T_C (without SHE) at a wide range of voltages.

$$cell = \begin{bmatrix} V_{ZTC_{(1,1)}} & \dots & V_{ZTC_{(1,7)}} \\ \vdots & \ddots & \vdots \\ V_{ZTC_{(7,1)}} & \dots & V_{ZTC_{(7,7)}} \end{bmatrix}$$

$$\text{NANDx2} = \begin{bmatrix} 0.53 & 0.53 & 0.52 & 0.52 & 0.51 & 0.50 & 0.49 \\ 0.53 & 0.53 & 0.53 & 0.52 & 0.51 & 0.50 & 0.49 \\ 0.53 & 0.53 & 0.53 & 0.53 & 0.51 & 0.51 & 0.50 \\ 0.54 & 0.54 & 0.53 & 0.53 & 0.53 & 0.51 & 0.50 \\ 0.54 & 0.54 & 0.54 & 0.53 & 0.53 & 0.53 & 0.53 \\ 0.54 & 0.54 & 0.54 & 0.54 & 0.53 & 0.53 & 0.53 \\ 0.55 & 0.54 & 0.54 & 0.54 & 0.54 & 0.54 & 0.53 \end{bmatrix}$$

The NANDx2 exhibits V_{ZTC} ranges between 0.55−0.49V with a majority of ZTC at 0.53V. Still, there is a clear trend indicating a dependency on both t_{slew} and C_{load}.

To highlight the variances in all V_{ZTC}, Fig. 4 shows the histogram of all simulation results of V_{ZTC}. Experiments cover all operating conditions for all cells (101 standard cells ×7 t_{slew} × 7 C_{load} = 4949 simulations and resulting V_{ZTC} values). The figure shows that the highest percentage of ZTC occurrence is at 0.54V, yet the span is still quite large from 0.49V to 0.55V. With such variance in V_{ZTC} within each cell and across cells, it is impossible to operate every cell in the circuit *exactly* at ZTC. As a result, a given circuit consists of subcircuits with different V_{ZTC}, since each cell (subcircuit) within has a different matrix. Therefore, finding overall V_{ZTC} of the circuit is challenging, as it is the weighted average of the V_{ZTC} of its subcircuits.

To distinguish V_{ZTC} from cells and chip, we refer to $V_{ZTC}(cell)$ and $V_{ZTC}(chip)$ from now on. $V_{ZTC}(chip)$ for the entire circuit is thus the weighted superposition of millions of $V_{ZTC}(cell)$ from all cell instances within it. However, this variance is minuscule, as we operate close to the ZTC for most cells as we explained later in Section V.

Please note, due to process variations, each transistor might have different characteristics. This results in a variation of ZTC of transistors. Our analysis shows that the variation of $V_{ZTC}(transistor)$ is small, and $V_{ZTC}(cell)$ is within the $V_{ZTC}(transistor)$ range (see Appendix B).

B. ZTC FOR LARGE CIRCUITS

Finding the ZTC voltage of a large circuit is challenging due to the different $V_{ZTC}(cell)$. Cells within the circuit should be examined for both delay and power under a set of conditions. With four dimensions t_{slew}, C_{load}, T_C and V_{dd} checking all these conditions is unfeasible due to simulation time. Therefore, we rely on the static timing analysis tools (STA) in order to find and then employ $V_{ZTC}(chip)$. Consequently, we operate with V_{dd} near ZTC (N-ZTC) of the individual cells. Our algorithm examines the circuit's delays at different T_Cs for a wide range of V_{dd}. When circuit's delays is identical (or within an acceptable delay variance ϵ) for a range of T_C, we found our $V_{ZTC}(chip)$. Our full approach for employing N-ZTC of a circuit is summarized in Algorithm 1.

First, the circuit's layout is designed after synthesizing the RTL of the circuit. With the layout available, signoff tool [18] creates best and worst-case corners for every voltage step based on given $T_C(low)$ and $T_C(high)$ temperatures (i.e., the highest and lowest T_C). T_C follows our results in Fig. 3, where $T_C = T_{chip} + \Delta T_C(\text{SHE})$. Note again that the worst-case is always the highest delay, not the highest temperature (e.g., in ITD region). The sign-off tool then estimates the circuit's delay t_{delay} at these T_C. By applying the worst-case approach and as the actual T_C is within the range of temperature, we guarantee functional operation of the circuit, i.e. our estimated guard band is able to protect the circuit against the temperature-induced delay shifts. The algorithm has to traverse all voltages within a suitable range (e.g., from $V_{ZTC}(pFinFET)$ to $V_{ZTC}(nFinFET)$) with the smallest possible step ($V_{step} = \alpha$), since we can not know in prior, where the V_{ZTC} might be. Iteratively, we reduce V_{dd} by a small step $\alpha = 0.01\text{V}$[1]. Each voltage, the analysis estimates at both high T_C ($T_C = T_{chip} + \Delta T_C(\text{SHE})$, see Fig. 3) and low T_C (without SHE, $T_C = T_{chip}$).

After that, our algorithm checks if we are near ZTC by comparing the t_{delay} at every V_{dd} for both worst and best corners. The accepted delay variance, in our work, is $\epsilon \leqslant 0.01$ ns (1% of our total $t_{delay}(CP) \approx 1$ ns).

C. SHE-AWARE STANDARD CELL LIBRARIES

Multi-Corner Multi-Mode (MCMM) are multiple executions of static timing analysis that used in the design of digital chips across all modes and corners concurrently. Available corners do not consider SHE. Hence, to analyze SHE of a circuit, it necessitates extending the available corners by creating *SHE-aware cell libraries*. In addition to higher temperatures (T_C), these cell libraries span a wide range of voltages to ensure ZTC is within our design space.

For this purpose, we characterize our own cell libraries by employing the SPICE netlists of combinational and sequential cells from the 7nm ASAP7 PDK [16]. The SHE-aware cell libraries are characterized considering the temperature used in the propagation delay simulations to the corresponding T_C under SHE. We tested three fin configurations: 1, 3, and 7 fins as shown in Fig. 3. This covers more than 90% of all transistors in the ASAP7 PDK, with the 3 fin transistor as the most occurring transistor in the ASAP7 cell library (40% is 3 fin). Considering worst-case operating, the opted to use the 7-fin SHE-induced degradation peak T_C as the temperature during characterization. This temperature is then entered in the library characterization tool to determine, via circuit simulations, power and delay of the standard cells under various t_{slew} and C_{load}. Delay and power of every cell are then stored within a lookup table in the *liberty* format.

We characterize the cell libraries for a set of voltages V_{dd} with the corresponding $T_C = T_{chip} + \Delta T_C(\text{SHE})$ (see Section III). To compare later on, we performed our entire process also without SHE ($T_C = T_{chip}$).

[1]On chip voltage regulators operate in 10mV intervals, see [19], [20].

Algorithm 1 Operating Near Zero-Temperature Coefficient (N-ZTC) Aiming Minuscule SHE-Induced Delay Variance

Require: *Voltage range, Voltage step α, channel ΔT_c list, SHE-aware libraries, chip layout, acceptable delay variance ϵ*
N-ZTC
Ensure: at V_{ZTC}
Set $V_{dd} = V_{Nominal}$ ▷ Start from nominal=0.7V
while ZTC not found **do**
  for Each ΔT_c in the list at V_{dd} (Fig. 3) **do**
    $T_C = T_{chip} + \Delta T_C$ ▷ T_C(SHE)
    Create Process corner at V_{dd} ▷ Using Voltus
    Set condition set Temperature = T_c
    Parasitics extraction ▷ Using Voltus
    STA Chip's delay analysis ▷ Using Tempus
    Report Delay $t_{delay}(T_c)$ ▷ Using Tempus
  end for
  $\Delta t_{delay} = t_{delay}(T_c(high)) - t_{delay}(T_c(low))$
  if $\Delta t_{delay} \leq \epsilon$ **then** ▷ acceptable delay variance ϵ
    ZTC found is True
  end if
  Update $V_{dd} = V_{dd} - \alpha$ ▷ update voltage
  Update T_c at V_{dd} ▷ update T_c list
end while
Report Power ▷ Using Voltus at ZTC point for all temperatures

V. EVALUATION

In the following, we present our approach following Algorithm 1. First, we describe our physical chip design of the processor. Then, we show $V_{ZTC}(cell)$ variance within the chip. Afterward, we determine N-ZTC of the entire chip ($V_{ZTC}(chip)$). Then, we compare our N-ZTC approach with traditional guard band in terms of performance and power. Lastly, we explain how multi-core systems are affected by SHE and N-ZTC in terms of performance, power, and energy.

A. PHYSICAL CHIP DESIGN

The physical design of a chip is the layout (full place and route) and post-synthesis optimization. Large chip designs likely feature higher T_C variance, due to more combinations

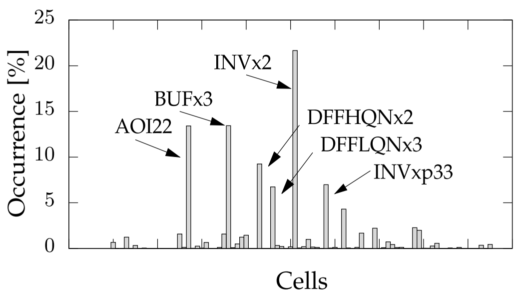

FIGURE 5. The used cells histogram within the OpenPiton chip layout (percentage of occurrences of each cell to the total number of cells).

of f_{sw} (switching frequency), t_{slew}, C_{load} for a wider variety of standard cells. Therefore, we target a relatively large circuit such as a full processor in order to maximize T_C variance. This work employs a full computing tile of the state-of-the-art OpenPiton processor, which is an open-source processor based on the OpenSPARC T1 core [21].

First, we synthesized the register-transfer level RTL of the processor using the baseline cell library from ASAP7 PDK [16] (i.e., at nominal voltage 0.7V) without SHE using a the Synopsys DC compiler [22]. Then, the design passed through place and route, including Power Delivery Network (PDN) design and optimization, using Cadence Innovus 7.1 [23]. Then, N-ZTC is determined based on post-layout simulations considering RC-parasitics and interconnects of the OpenPiton chip using the on-chip variation feature to consider their impacts on delay and power. Using the chip's layout within EDA tools, not solely the synthesized netlist (misses important information like RC-parasitics), allows us to accurately perform SHE analysis in different thermal regions. Since these tools can handle complex designs, we can employ N-ZTC regardless of the chip's size.

B. ZTC VARIANCE WITHIN OUR PROCESSOR

The designed chip consists of 448,668 different cells. The synthesis tool used 86 to build the circuit out of the available 101 standard cells in the PDK. Fig. 5 shows the histogram of the instantiated cells within the chip.

Selecting $V_{dd} = 0.54$V, as the major occurring V_{ZTC} from Fig. 4 would result in lots of cells operate exactly at their $V_{ZTC}(cell)$, some cells are in ITD and the remaining in PTD. Therefore, when operating at $V_{ZTC}(chip)$ it is a compromise and the cells are distributed over all three thermal regions.

To grasp the variations in V_{ZTC} we use the standard deviation σ in the used cells. We estimate σ of V_{ZTC}, defined in Eq. (1), for every operating condition (e.g., $V_{ZTC_{(1,1)}}$) across all cells in the OpenPiton processor.

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(V_{ZTCi} - \overline{V_{ZTC}})^2} \quad (1)$$

where σ is the standard deviation, N is the number of operating conditions, and $\overline{V_{ZTC}}$ is the arithmetic mean across all V_{ZTC} under the same t_{slew} and C_{load}.

$$\sigma = \begin{bmatrix} \sigma_{(1,1)} & \cdots & \sigma_{(1,7)} \\ \vdots & \ddots & \vdots \\ \sigma_{(7,1)} & \cdots & \sigma_{(7,7)} \end{bmatrix}$$

Therefore, $\sigma = 0$ indicates that only a single $V_{ZTC} = \overline{V_{ZTC}}$ exists across the cells, i.e. all cells have identical V_{ZTC} under given t_{slew} and C_{load}. Vice versa, $\sigma > 0$ indicates different V_{ZTC} distinct from the mean $\overline{V_{ZTC}}$. Spanning the operating conditions, we observe that the majority of cells operate at V_{ZTC} of $\overline{V_{ZTC}} \approx 0.53$V contrary to the most occurring voltage of 0.54V (38%) in Fig. 4. However, 0.53V is the second-most occurring voltage with 33% of all values in Fig. 4. This small difference results from the selection of cells and their surroundings stemming from the synthesis tool. Results of σ are summarized in the following matrix:

$$\sigma = \begin{bmatrix} 0 & 0 & 0.007 & 0.018 & 0.021 & 0.027 & 0.033 \\ 0 & 0 & 0 & 0.005 & 0.01 & 0.016 & 0.027 \\ 0 & 0.04 & 0 & 0 & 0.06 & 0.013 & 0.025 \\ 0.18 & 0.15 & 0.1 & 0 & 0 & 0.007 & 0.013 \\ 0.21 & 0.17 & 0.13 & 0.09 & 0 & 0 & 0 \\ 0.24 & 0.2 & 0.18 & 0.1 & 0.01 & 0 & 0 \\ 0.31 & 0.27 & 0.23 & 0.19 & 0.14 & 0.06 & 0 \end{bmatrix}$$

This highlights how under the same t_{slew} and c_{load}, different cells exhibit different V_{ZTC}. Therefore, it is impossible to operate each cell *exactly* at its V_{ZTC}. Instead, a compromise must be found. Instead of finding V_{ZTC} for every subcircuit (standard cells in our case) to find $V_{ZTC}(chip)$, we directly estimate $V_{ZTC}(chip)$ as discussed in Section IV-B.

C. DETERMINING ZTC OF THE OpenPiton PROCESSOR

To determine $V_{ZTC}(chip)$, we implement Algorithm 1. For each iteration, V_{dd} is reduced by the smallest possible step α (e.g., 0.01V) and then t_{delay} of the chip is examined with SHE ($T_C(high)$) and without SHE ($T_C(low)$) using Signoff tools based on our SHE-aware cell libraries. The chip's delay results (t_{delay}) of low and high T_C over voltage converge towards $V_{ZTC}(chip)$. Since our voltage range is large enough, we must cross from ITD region to the PTD region and thus pass ZTC. Hence, Algorithm 1 must terminate with $V_{ZTC}(chip)$.

Lowering the supply voltage reduces $\Delta T_C(SHE)$ and thus $T_C(high) = T_{chip} + \Delta T_C(SHE)$. Therefore, we do not solely gain performance due to the lower (or even zero) timing guard band, but also *lower* the temperature T_C. This is important as T_C stimulates other reliability phenomena like aging effects [10], [13] and thus lowering T_C lowers aging, in term of reducing the guard band to protect against aging. Please note, that aging-induced degradations are reducing much faster than the resilience against aging. Hence operating at V_{ztc}, reduces the required aging-guard band [24], [25]. Additionally, our delay and power estimations are based on the *variable* temperature with voltage changes.

Fig. 6 shows t_{delay} of the processor's chip with SHE ($T_c(high)$) and without SHE ($T_c(low)$) over a wide range

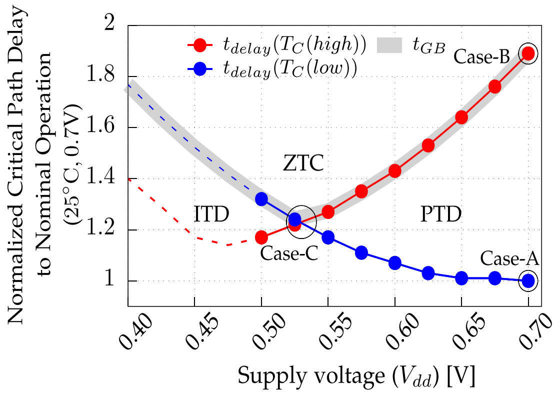

FIGURE 6. The processor's delay changes with T_C due to SHE normalized to the base operating condition (25°C, 0.7V). Both delays (i.e., T_C(low) and T_C(high)) are matched near V_{dd}≈0.53V. The delay of T_C(high) is expected to increase after dependencies changed as predicted in dashed lines. Guard bands are always the worst-case delay, regardless if it occurs at high or low T_C. The figure shows the possible operating point cases; Case-A: nominal without SHE, Case-B: traditional guard band, and Case-C: N-ZTC.

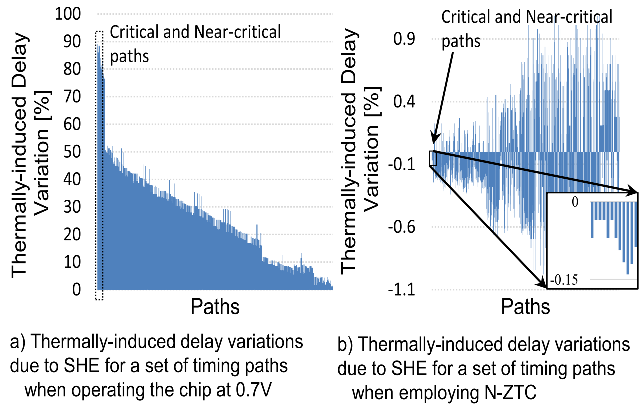

FIGURE 7. SHE-induced delay degradations of a set of paths within the chip. (a) variances due to SHE-induced delay degradation when operating at the nominal voltage (0.7V) without SHE. (b) variances due to SHE-induced delay degradation employing N-ZTC (V_{dd} = 0.53V).

of supply voltage V_{dd} where the thermal regions can be clearly identified (PTD, ZTC, and ITD). The delay is normalized to the nominal operating condition (V_{dd} = $0.7V$, 25°C). *Guard band follows the worst-case delay as shown in the same figure with the gray curve*. Hence, t_{clk} is always max($t_{delay}(T_C(low))$, $t_{delay}(T_C(high))$). The delay of both curves is expected to increase after a certain point as shown in the dashed line. This happens when $T_C(low) = T_{chip} = T_C(high)$ since at low voltages $\Delta T_C(SHE)$ tends to zero or when the thermal dependence of the chips delay becomes weaker than the voltage dependence. As shown, $V_{ZTC}(chip)$ occurs near 0.53V. This $V_{ZTC}(chip)$ is closer to nominal V_{dd} than previously reported [6], [7] due to the smaller technology. This makes operating at ZTC more feasible, as the induced performance degradation is smaller if $\Delta V = V_{nominal} - V_{ZTC}$ is small. However, the $V_{ZTC}(chip)$ is intended to be a chip specific and will differ from one chip to another.

D. TRADITIONAL GUARD BANDS FOR SHE MITIGATION

The guard band to mitigate SHE-induced delay degradation in our processor is shown in grey in Fig. 6. SHE-induced delay degradation at nominal V_{dd} is high where $t_{GB} > 90\% \cdot t_{delay}(CP)$ and hence the circuit operate at a much higher delay (i.e., $t_{clk} = t_{delay}(CP) + t_{GB}$). The guard band t_{GB} reduces when V_{dd} reduces (starting in PTD) until it reaches ZTC. Reducing voltage below $V_{ZTC}(chip)$, in the ITD region, increases the worst delay *(now low instead of high T_C due ITD)* again.

Operating the chip at $V_{ZTC}(chip)$ is a compromise. The delay of the chip is determined by the variances in the critical timing paths. The final delays of the critical paths experience minuscule thermally induced delay variance. Our investigation shows a delay variance of <0.1% in critical and near-critical timing paths. This is due to the acceptable error ϵ (i.e., tolerance factor) that employed in our algorithm as we have 10mV voltage steps and thus might miss the perfect V_{ZTC}. With this small delay variance, the required guard band is also small where t_{GB}<0.02ns (i.e., near-zero guard band). However, non-critical paths exhibit larger thermally induced delay variance. Non-critical paths are by definition not critical, i.e., do not determine the timing of the entire chip. This is by design, as Algorithm 1 used timing analysis of the entire chip to determine $V_{ZTC}(chip)$. Our approach considers near-critical paths becoming critical and always finds the path with the worst delay to determine $t_{delay}(chip)$. However, all the other paths might still feature a negligible variance which has no impact on the overall chip timing.

To illustrate the delay variances within the timing paths, we examined a sample set (we can not show millions of paths), that covers a wide range of t_{delay} from timing paths (i.e., critical and non-critical paths) within the chip. Fig. 7a shows SHE-induced delay variances of the chip operating at nominal voltage (V_{dd} = 0.7V) where all paths are prolonged in their delay, as all cells operate in PTD and T_C is elevated. Comparing Fig. 7a to Fig. 7b, which shows SHE-induced delay variances of the chip employing N-ZTC ($V_{dd} = V_{ZTC}(chip)$ = 0.53V), we can clearly see that thermally induced delay variance in our approach is <0.1%. This is expected and thus our approach worked fine. At the same time, delay variances in non-critical paths are larger (i.e., $\sigma(t_{delay})<\pm1\%$). This is not an issue, as they will never become critical and thus cannot introduce timing violations. Nevertheless, the designer should be aware that we only minimize the variance here. Still, note that original delay variance was $\sigma(t_{delay})>90\%$ and now became $\sigma(t_{delay})<\pm1\%$, so also the non-critical paths received a vast improvement in terms of delay variance.

Comparison Between Nominal Operation, Traditional t_{GB} and N-ZTC: We compare here the three possible operating points: case-A: Baseline at nominal voltage without SHE ($T_C(low)$), case-B: Traditional guard band at nominal voltage with SHE ($T_C(high)$), and case-C: N-ZTC operation (lower V_{dd} and any T_C). All cases are shown in Fig. 6.

Case-C (N-ZTC) does not reach the performance of case-A without any guard band. This is expected, as case-A would immediately exhibit timing violations if the temperature would increase above nominal temperature (e.g., room temperature). Instead, a delay degradation of 25% is observed due to the lower V_{dd} when moving from nominal V_{dd} to $V_{ZTC}(chip)$. However, we can observe a 65% performance improvement due to a reduction of t_{GB} compared to Case-B (traditional guard band). In terms of power, N-ZTC results in less leakage power compared to Case-B and Case-A due to the reduced supply voltage, despite the elevated leakage from operating at high temperatures. The results are summarized in Table 1 in comparison with the theoretical baseline case-A.

TABLE 1. Comparison between the three possible operating points: Baseline, traditional guard band, and N-ZTC. Results are compared to case-A.

Case	V_{dd}	GB	Delay increase	Leakage Power	Freq.	Reliable
A(Baseline)	0.7V	No	0 [%]	100 [%]	1.77GHz	No
B(Traditional)	0.7V	Large	91 [%]	600 [%]	0.95GHz	Yes
C(N-ZTC)	0.53V	Near-zero	25 [%]	39 [%]	1.45GHz	Yes

VI. MULTI-CORE ANALYSIS

This section evaluates if employing N-ZTC is beneficial to the computing system as a whole. Previously, we directly linked delay to performance, i.e. minimizing guard bands increases performance while reducing the voltage to $V_{ZTC}(chip)$ (i.e., aiming N-ZTC) requires scaling down the frequency and therefore reducing the performance. However, the system performance (e.g., makespan or throughput of an application) differs from the circuit performance (e.g., cycles per second). This section evaluates if there is an overall gain in system performance. Next to evaluating performance, we also evaluate the impact of N-ZTC ($V_{dd} = V_{ZTC}(chip)$) on energy consumption (e.g., battery life). From the previous section, it is clear that lowering V_{dd} reduces leakage power. However, with execution time rising and power dropping, energy (power delay product) might increase or decrease. This section evaluates if operating at $V_{ZTC}(chip)$ saves energy, in a multi-core system.

A. EXPERIMENTAL SETUP

We simulate a multi-core with four out-of-order cores modeling the *Gainestown* micro-architecture. Each of the cores is associated with private L1-I and L2-D caches with 32 KB each, as well as a private 256 KB L2 cache. Additionally, the multi-core contains an 8 MB shared L3 cache.

The multi-core is modeled to be implemented with the same 7nm PDK as used for OpenPiton design (see Section V-A). We use the *Sniper* [26] many-core simulator, which allows multi-threaded simulation with full modeling of shared resource contention. *McPAT* [27] is used to estimate the power and energy consumption of the simulated multi-core. We execute applications from the *PARSEC* benchmark suite [28] with *simlarge* inputs. These applications cover compute-bound applications like *blackscholes* as well as memory-bound applications like *canneal*.

Because *McPAT* does not support 7nm FinFET, which is our target technology, we scale the power values obtained from estimations performed with 45nm using low-power devices (smallest supported technology). In order to scale the power from 45nm to 7nm, we implement the *OpenPiton* SoC using both a 45nm Bulk CMOS [29] and 7nm conventional FinFET [16] to obtain scaling factors for dynamic and leakage power. These implementations follow the same approach as described in Section V.

B. COSTS AND BENEFITS FROM N-ZTC

Cases: We explore again the previous three cases shown in Fig. 6 and described in Table 1. Case-A is the baseline design, i.e., SHE-unaware. No guard bands are applied, which allows operating the multi-core at its peak frequency of 1.77 GHz. *This case features SHE-induced timing violations as it ignores the impact of SHE on the delay. While unreliable, this case acts as a baseline to see what theoretical performance would be achievable if SHE or thermal degradation in general would not be an issue.* Case-B applies traditional guard bands. It accounts for delay increases due to SHE and therefore adds a timing guarband to its clock frequency, resulting in a lower frequency (0.95 GHz). Case-C employing N-ZTC, which is operation at $V_{ZTC}(chip) = 0.53$ V. *Here, near-zero timing guard bands for temperature-induced degradation (e.g., SHE) are needed (<0.1%).* Yet, V_{dd} is below nominal and as such the same clock frequency cannot be maintained. So instead of a guard band lowering the frequency, now it is the lower supply voltage, which reduces 1.77 to 1.45 GHz. As can be noticed, this is faster than traditional guard banding in terms of circuit performance.

Usecase: We execute four-threaded *PARSEC* applications to fully utilize the studied multi-core and operate the cores at the voltage and frequency defined by each case. We record the benchmark execution time as a measure for system performance and the corresponding energy consumption.

Execution time: Fig. 8a shows the execution time for different applications with the three cases. Results are normalized to case-A. System performance of case-A is our theoretical value and is much faster than the reduced frequency in case-B and slightly faster than case-C. However, the operating frequency does not represent system performance. What matters is the actual runtime of applications on our processor, i.e. how long a given task takes. Importantly, applications suffer unequally from reduced frequencies. While the performance of compute-bound applications like *blackscholes* scales almost linearly with the CPU frequency, the performance of memory-bound applications like *canneal* depends strongly on the L3 and DRAM frequency, which is unaltered by operating at V_{ZTC}. In summary, N-ZTC exhibits better system performance for all applications compared to traditional SHE guard band and is comparable in system

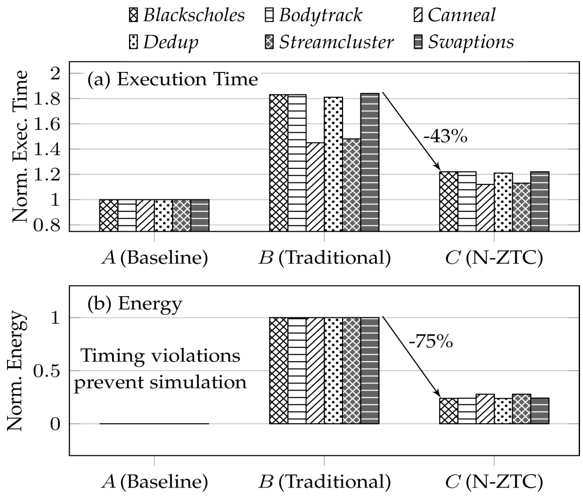

FIGURE 8. Execution time and energy of the three cases. Baseline case does not employ SHE guard bands and therefore does not allow reliable execution.

performance to the theoretical upper bound for memory-bound applications.

Energy: Fig. 8b presents the energy consumed for the execution of different applications. The results are normalized to case-B, as it consumes the most energy. Case-B uses the same voltage as case-A but at a lower frequency due to the guard bands. This means, that it takes the longest execution time. Yet, the important question if case-B consumes more or less energy than case-A which could not be answered, since timing violations prevented a simulation at elevated T_C in case-A. We have to elevate T_C to consider the leakage increase due to temperature and this results in timing violations. Therefore, case-A is unrealistic since it causes timing violations, thus we neglect its results.

VII. CONCLUSION

SHE-induced delay degradation, traditionally, can be mitigated by employing a *large timing guard band* to guarantee operation without errors. This work exploited operating near Zero-Temperature Coefficient (N-ZTC) to minimize the impact of SHE on the circuit's delay and eliminate the need for large guard bands. We presented our algorithm aiming to accurately locate the proper voltage to operate at $V_{ZTC}(chip)$. Results show that near-zero guard band is still required when operating N-ZTC. Simulations of both circuit and system levels show a significant enhancements in term of performance (up to 65%) and leakage power (up to 94%) when employing N-ZTC in comparison with traditional guard band technique. Multi-core simulations show 43% lower performance loss and 75% lower energy on average when comparing N-TZC operation with traditional guard banding at nominal V_{DD}.

APPENDIX A BACKGROUND

Here, we explain some important background details.

A. FIGURATIVE IMPACT OF SHE ON TRANSISTORS

Temperature affects two key parameters in a transistor: threshold voltage (V_{th}) and carrier mobility (μ) [2]. In its simplest form, both parameters can be modeled as functions of temperature according to [4]:

$$\mu(T_C) = \mu(T_{ambient})(\frac{T_{ambient}}{T_C})^m \quad (2)$$

$$V_{th}(T_C) = V_{th}(T_{ambient}) - k(T_C - T_{ambient}) \quad (3)$$

where $T_{ambient}$ is the room temperature in Kelvin, m and k are positive constants, and T_C is channel temperature. These models show that V_{th} scales linearly with an increase in T_C, while μ scales with a power law. This explains the origin behind the thermal regions.

B. TEMPERATURE MODELING OF TRANSISTORS

While for large transistors, Eq. (3) and Eq. (2) from 2001 [4] were fine. Nano-scale transistors have various additional dependencies, which must be considered. The temperature models $V_{th}(T_C)$ and $\mu(T_C)$, as well as the resulting $I_D(T_C)$, need to be more sophisticated to accurately predict transistor behavior and match reported experimental data. V_{th} temperature dependency:

$$V_{th} = V_{th0} + \Delta V_{th}, all^2 \quad (4)$$

$$V_{th0} = \frac{kT}{q} \cdot ln\left[\frac{C_{ox}\frac{kT}{q} \cdot (C_{ox}\frac{kT}{q} + 2Q_{bulk} + 5C_{si}\frac{kT}{q})}{2q \cdot n_i \cdot \epsilon_{sub} \cdot \frac{kT}{q}}\right] + V_{fb} + \phi_B + \Delta V_{th,QM} + \frac{kT}{q} + q_{bs} \quad (5)$$

Where the following parameters are temperature dependent (i.e., feature the term) "$\frac{kT}{q}$": C_{ox} is the oxide capacitance, C_{si} is the body capacitance, Q_{bulk} is the fixed depletion charge, $\Delta V_{th,QM}$ is the surface potential considering quantum mechanical effect, k is boltzmann constant, q is the electronic charge, n_i is the intrinsic carrier concentration, T is the temperature, ϵ_{sub} is the dielectric constant. V_{fb} is the flatband voltage, ϕ_B is the body-effect voltage parameter, q_{bs} is the body doping. Note the frequent occurrence of temperature terms "$\frac{kT}{q}$", which highlights the actual complexity of taking elevated T_C into account.

C. THERMAL REGIONS

Normally, the circuit's delay increases when the temperature increases. However, lowering the supply voltage will change this dependence. A decrease in V_{th} due to temperature rise increases I_D by $\Delta I_D(V_{th})$, while a decrease in μ decreases I_D by a different amount $\Delta I_D(\mu)$. Therefore, V_{th} and μ have opposing effects on I_D. As the thermal dependencies (Eq. (2) and Eq. (3)) are different in strength, lowering supply voltage (V_{dd}) changes the strength of the two opposing forces drawing on I_D. Hence, three regions emerge: Positive-Temperature Dependence (PTD), Zero-Temperature Coefficient (ZTC) and an Inverse-Temperature Dependence (ITD) as shown in Fig. 9. In these three regions, I_D falls, stays exactly the same or rises with increasing T_C,

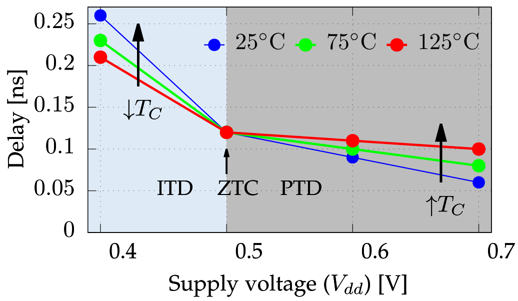

FIGURE 9. Definition of the thermal regions when operating a Ring Oscillator (RO) circuit, consisting of 13 inverters designed at 7nm technology [16], at different voltages and three different temperatures where three regions emerge: Positive-Temperature Dependence (PTD), Zero-Temperature Coefficient (ZTC) and Inverse-Temperature Dependence (ITD). Please note that RO's circuit is absolutely uniform as we simulated identical cells, therefore, ZTC is *identical* for all cells and no thermal variance is exhibited at ZTC.

depending if $\Delta I_D(\mu)$ is larger or smaller than $\Delta I_D(V_{th})$. Following the proposed methodology in this paper, we have tested an RO circuit for ZTC. Fig. 9 shows the delay of the critical path $t_{delay}(CP)$ of a ring oscillator (RO), consisting of 13 inverters designed at 7nm technology [16], operating at three T_Cs over voltage. Delay values $t_{delay}(CP)$ start to converge in the PTD region with V_{dd} decreases. This trend remains until all $t_{delay}(CP)$ values meet at ZTC. Continuing over V_{dd} decreases, $t_{delay}(CP)$ values start to diverge again in the opposite direction in ITD. At ZTC ($V_{ZTC} = 0.5$V in this example), $\Delta I_D(\mu) = \Delta I_D(V_{th})$ and thus, transistors (and thus the circuit) do not exhibit any thermal variance due to the compensation of beneficial ΔV_{th} with detrimental $\Delta\mu$. Please note that RO's circuit is absolutely uniform ignoring local variation, i.e., all subcircuits are identical inverter standard cells. Therefore, ZTC is *identical* for all subcircuits and no thermal variance is exhibited when operating at ZTC.

D. TIMING GUARD BAND

Timing guard band is typically employed in order to tolerate any runtime degradation in the delay of the circuit. Traditionally, designers employ the worst-case timing scenario to overcome SHE-induced delay degradation (i.e., delay increases). Timing guard band (t_{GB}) is a time added on top of the maximum delay of a circuit (i.e., critical path delay $t_{delay}(CP)$) to overcome delay degradations. This corresponds to a timing slack applied to the clock period shown in Eq. (6).

$$\begin{aligned} t_{clk} &= t_{delay}(CP) + t_{GB} \\ t_{GB} &= \Delta t_{delay}(CP) \end{aligned} \tag{6}$$

where $t_{delay}(CP)$ is the nominal propagation delay of the critical path in the circuit, t_{GB} is the deliberate timing margin added to tolerate degradation (e.g., shifts in path delay $\Delta t_{delay}(CP)$) and t_{clk} the clock period. Larger $\Delta t_{delay}(CP)$ necessitates longer t_{GB} and thus longer t_{clk}, reducing f_{clk} and thus the performance of the circuit. Therefore, t_{GB} must be minimized in order to keep performance as high as possible.

Nevertheless, t_{GB} tolerates degradations regardless if they occur during higher or low temperatures. It does not matter if t_{delay} starts to shift due to a high or low temperature from its nominal value. The guard band t_{GB} always follows worst-case timing. In ITD this means t_{delay} at low T_C, while in PTD this means t_{delay} at high T_C.

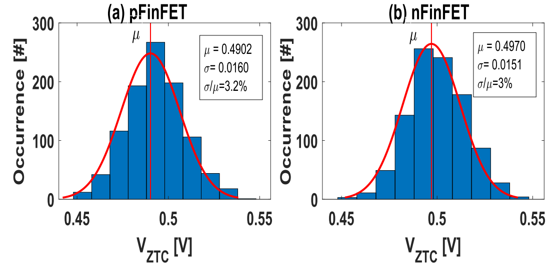

FIGURE 10. The histogram of ZTC of a) pFinFET and b)nFinFET transistors under process variations. V_{ZTC} values for both transistor types are distributed within a small range [0.45V - 0.55V].

APPENDIX B
ZTC OF TRANSISTORS UNDER PROCESS VARIATIONS

Due to process variations, each transistor within the circuit could have different characteristics. This results in a variation of ZTC of transistors. To demonstrate such variation, we simulate 1000 different nFinFET and 1000 different pFinFET transistors (i.e., different length, width, etc.) using HSPICE. The actual variability data are taken from [30], [31] for Intel 14nm FinFET technology. We study the variations for T_C high and low for a large range of voltages [0.2V-0.7V] with 10mV steps (see Algorithm 1). To determine ZTC of a transistor, we examine I_d of the transistor at high and low T_C. The voltage that shows no difference in I_d (because the propagation delay of the transistor is function of I_d) is therefore our ZTC. Results show that V_{ZTC} values for both transistor types are distributed within a small range [0.45V - 0.55V] as demonstrated in Fig. 10. Importantly, by design, V_{ZTC} *of a chip must be located within this small range*.

REFERENCES

[1] J. Hwan Choi, J. Murthy, and K. Roy, "The effect of process variation on device temperature in finFET circuits," in *Proc. IEEE/ACM Int. Conf. Comput.-Aided Design*, Nov. 2007, pp. 747–751, doi: 10.1109/ICCAD.2007.4397355.

[2] D. Wolpert and P. Ampadu, "Temperature effects in semiconductors," in *Managing Temperature Effects in Nanoscale Adaptive Systems*. New York, NY, USA: Springer, 2012, pp. 15–33.

[3] Y. Tsividis and C. McAndrew, *Operation and Modeling of the MOS Transistor* (Oxford Series in Electrical and Computer Engineering), 3rd ed. New York, NY, USA: Oxford Univ. Press, 2011. [Online]. Available: https://cds.cern.ch/record/1546736

[4] K. Kanda, K. Nose, H. Kawaguchi, and T. Sakurai, "Design impact of positive temperature dependence on drain current in sub-1-V CMOS VLSIs," *IEEE J. Solid-State Circuits*, vol. 36, no. 10, pp. 1559–1564, Oct. 2001, doi: 10.1109/4.953485.

[5] A. Calimera, R. I. Bahar, E. Macii, and M. Poncino, "Temperature-insensitive dual-V_{th} synthesis for nanometer CMOS technologies under inverse temperature dependence," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 18, no. 11, pp. 1608–1620, Nov. 2010, doi: 10.1109/TVLSI.2009.2025884.

[6] D. Bol, C. Hocquet, D. Flandre, and J.-D. Legat, "The detrimental impact of negative celsius temperature on ultra-low-voltage CMOS logic," in *Proc. Eur. Solid-State Circuits Conf. (ESSCIRC)*, Sep. 2010, doi: 10.1109/ESSCIRC.2010.5619758.

[7] M. Cho, M. Khellah, K. Chae, K. Ahmed, J. Tschanz, and S. Mukhopadhyay, "Characterization of inverse temperature dependence in logic circuits," in *Proc. IEEE Custom Integr. Circuits Conf.*, Sep. 2012, pp. 1–4, doi: 10.1109/CICC.2012.6330659.

[8] E. Pop, R. Dutton, and K. Goodson, "Thermal analysis of ultra-thin body device scaling [SOI and FinFet devices]," in *IEDM Tech. Dig.*, Dec. 2003, pp. 36.6.1–36.6.4, doi: 10.1109/IEDM.2003.1269420.

[9] 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, doi: 10.1109/16.259622.

[10] W. Ahn, S. H. Shin, C. Jiang, H. Jiang, M. A. Wahab, and M. A. Alam, "Integrated modeling of Self-heating of confined geometry (FinFET, NWFET, and NSHFET) transistors and its implications for the reliability of sub-20 nm modern integrated circuits," *Microelectron. Rel. (MR)*, vol. 81, pp. 262–273, Feb. 2018, doi: 10.1016/j.microrel.2017.12.034.

[11] D. Jang, E. Bury, R. Ritzenthaler, M. G. Bardon, T. Chiarella, K. Miyaguchi, P. Raghavan, A. Mocuta, G. Groeseneken, A. Mercha, D. Verkest, and A. Thean, "Self-heating on bulk FinFET from 14 nm down to 7 nm node," in *IEDM Tech. Dig.*, Dec. 2015, pp. 11.6.1–11.6.4, doi: 10.1109/IEDM.2015.7409678.

[12] H. Jiang, S. Shin, X. Liu, X. Zhang, and M. A. Alam, "The impact of self-heating on HCI reliability in high-performance digital circuits," *IEEE Electron Device Lett.*, vol. 38, no. 4, pp. 430–433, Apr. 2017, doi: 10.1109/LED.2017.2674658.

[13] J. Henkel and N. Dutt, *Dependable Embedded Systems*. Cham, Switzerland: Springer, 2021, doi: 10.1007/978-3-030-52017-5.

[14] J. P. Duarte, S. Khandelwal, A. Medury, C. Hu, P. Kushwaha, H. Agarwal, A. Dasgupta, and Y. S. Chauhan, "BSIM-CMG: Standard FinFET compact model for advanced circuit design," in *Proc. Conf. 41st Eur. Solid-State Circuits Conf. (ESSCIRC)*, Sep. 2015, doi: 10.1109/ESSCIRC.2015.7313862.

[15] U. S. Kumar and V. R. Rao, "A thermal-aware device design considerations for nanoscale SOI and bulk FinFETs," *IEEE Trans. Electron Devices*, vol. 63, no. 1, pp. 280–287, Jan. 2016.

[16] L. T. Clark, V. Vashishtha, L. Shifren, A. Gujja, S. Sinha, B. Cline, C. Ramamurthy, and G. Yeric, "ASAP7: A 7-nm finFET predictive process design kit," *Microelectron. J.*, vol. 53, pp. 105–115, Jul. 2016, doi: 10.1016/j.mejo.2016.04.006.

[17] Synopsys. *SAED Cell Library*. Accessed: Dec. 2020. [Online]. Available: http://web.engr.oregonstate.edu/~traylor/ece474/reading/saed_cell_lib_rev1_4_20_1.pdf

[18] (2018). *Voltus IC Power Integrity Solution*. [Online]. Available: https://www.cadence.com/content/cadence-www/global/en_US/home/tools/digital-design-and-signoff/silicon-signoff/voltus-ic-power-integrity-solution.html

[19] B. Keller, M. Cochet, B. Zimmer, Y. Lee, M. Blagojevic, J. Kwak, A. Puggelli, S. Bailey, P.-F. Chiu, P. Dabbelt, C. Schmidt, E. Alon, K. Asanović, and B. Nikolić, "Sub-microsecond adaptive voltage scaling in a 28 nm FD-SOI processor SoC," in *Proc. ESSCIRC Conf. 42nd Eur. Solid-State Circuits Conf.*, Sep. 2016, pp. 269–272, doi: 10.1109/ESSCIRC.2016.7598294.

[20] Z. Kamal, Q. Hassan, and Z. Mouhcine, "Full on chip capacitance PMOS low dropout voltage regulator," in *Proc. Int. Conf. Multimedia Comput. Syst.*, Apr. 2011, pp. 1–4, doi: 10.1109/ICMCS.2011.5945660.

[21] J. Balkind, M. McKeown, Y. Fu, T. Nguyen, Y. Zhou, A. Lavrov, M. Shahrad, A. Fuchs, S. Payne, X. Liang, M. Matl, and D. Wentzlaff, "OpenPiton: An open source manycore research framework," in *Proc. Int. Conf. Architectural Support Program. Lang. Operating Syst. (ASPLOS)*, 2016, pp. 217–232, doi: 10.1145/2872362.2872414.

[22] (2018). *Synopsys EDA Tool Flows*. [Online]. Available: https://www.synopsys.com/

[23] (2018). *Cadence EDA Tool Flows*. [Online]. Available: https://www.cadence.com/

[24] V. M. van Santen, H. Amrouch, N. Parihar, S. Mahapatra, and J. Henkel, "Aging-aware voltage scaling," in *Proc. Design, Autom. Test Eur. Conf. Exhib. (DATE)*, 2016, pp. 576–581.

[25] V. Santen, J. Martin-Martinez, H. Amrouch, M. M. Nafria, and J. Henkel, "Reliability in super- and near-threshold computing: A unified model of RTN, BTI, and PV," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 65, no. 1, pp. 293–306, Jan. 2017.

[26] T. E. Carlson, W. Heirman, and L. Eeckhout, "Sniper: Exploring the level of abstraction for scalable and accurate parallel multi-core simulation," in *Proc. Conf. High Perform. Comput., Netw., Storage Anal. (SC)*, 2011, pp. 1–12, doi: 10.1145/2063384.2063454.

[27] S. Li, J. H. Ahn, R. D. Strong, J. B. Brockman, D. M. Tullsen, and N. P. Jouppi, "The McPAT framework for multicore and manycore architectures: Simultaneously modeling power, area, and timing," *ACM Trans. Archit. Code Optim.*, vol. 10, no. 1, pp. 1–29, Apr. 2013, doi: 10.1145/2445572.2445577.

[28] C. Bienia, S. R. Kumar, J. Singh, and K. Li, "The PARSEC benchmark suite: Characterization and architectural implications," in *Proc. Int. Conf. Parallel Archit. Compilation Techn. (PACT)*, 2008, pp. 72–81.

[29] *Nangate, Open Cell Library*. Accessed: Dec. 2020. [Online]. Available: https://silvaco.com/services/library-design/

[30] H. Amrouch, G. Pahwa, A. D. Gaidhane, C. K. Dabhi, F. Klemme, O. Prakash, and Y. S. Chauhan, "Impact of variability on processor performance in negative capacitance finfet technology," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 67, no. 9, pp. 3127–3137, Sep. 2020.

[31] S. Natarajan *et al.*, "A 14 nm logic technology featuring 2^{nd}-generation FinFET, air-gapped interconnects, self-aligned double patterning and a 0.0588 μm^2 SRAM cell size," in *IEDM Tech. Dig.*, Dec. 2014, pp. 3.7.1–3.7.3, doi: 10.1109/IEDM.2014.7046976.

SAMI SALAMIN (Student Member, IEEE) received the B.Sc. degree in computer systems engineering and the M.Sc. degree (Hons.) from Palestine Polytechnic University, Hebron, Palestine, in 2005 and 2012, respectively. He is currently pursuing the Ph.D. degree with the Chair of Embedded Systems (CES), Karlsruhe Institute of Technology (KIT).

His research interests include reliable circuit design and analysis, emerging technology, low power design, and machine learning in the IoT.

VICTOR M. VAN SANTEN (Member, IEEE) received the Dipl.-Inf.(M.Sc.) degree in computer science from the Karlsruhe Institute of Technology (KIT), in 2014. He is currently a Researcher with the Chair of Semiconductor Test and Reliability (STAR), University of Stuttgart. His research interests include reliable circuit design and aging phenomena from the defect to the micro-architecture level.

MARTIN RAPP (Graduate Student Member, IEEE) received the B.Sc. and M.Sc. degrees (Hons.) in computer science from the Karlsruhe Institute of Technology, in 2014 and 2016, respectively. He is currently pursuing the Ph.D. degree under the supervision of Dr. Jörg Henkel. His research interest includes resource management for many-core processors with a focus on thermal management, where he is looking into how machine-learning-based techniques can support run-time management.

JÖRG HENKEL (Fellow, IEEE) received the Diploma and Ph.D. *(summa cum laude)* degrees from the Technical University of Braunschweig. He was a Research Staff Member with NEC Laboratories, Princeton, NJ, USA. He is currently the Chair Professor of Embedded Systems with the Karlsruhe Institute of Technology. He coordinates the DFG Program SPP 1500 Dependable Embedded Systems and is a Site Coordinator of the DFG TR89 Collaborative Research Center on Invasive Computing. His research work is focused on co-design for embedded hardware/software systems with respect to power, thermal, and reliability aspects. He has received six best paper awards throughout his career from, among others, ICCAD, ESWeek, and DATE. For two consecutive terms he served as the Editor-in-Chief for the *ACM Transactions on Embedded Computing Systems*. He is currently the Editor-in-Chief of the *IEEE Design&Test* Magazine and is/has been an Associate Editor of major ACM and IEEE journals. He has led several conferences as a general chair, including ICCAD and ESWeek and serves as a steering committee chair/member for leading conferences and journals for embedded and cyber-physical systems. He is the Chairman of the IEEE Computer Society and Germany Chapter.

HUSSAM AMROUCH (Member, IEEE) received the Ph.D. degree *(summa cum laude)* from KIT, in 2015. He is currently a Junior Professor with the Chair of Semiconductor Test and Reliability (STAR), Faculty of Computer Science, Electrical Engineering and Information Technology, University of Stuttgart, and a Research Group Leader with the Karlsruhe Institute of Technology (KIT), Germany. He has more than 100 publications in multidisciplinary research areas across the entire computing stack, starting from semiconductor physics to circuit design all the way up to computer-aided design and computer architecture. His main research interests include design for reliability and testing from device physics to systems, machine learning, security, approximate computing, and emerging technologies with a special focus on ferroelectric devices. He holds seven HiPEAC paper awards and three best paper nominations at top EDA conferences, such as DAC'16, DAC'17, and DATE'17, for his work on reliability. He also serves as an Associate Editor of *Integration*, the VLSI Journal. He has served in the technical program committees of many major EDA conferences, such as DAC, ASP-DAC, and ICCAD and as a reviewer for many top journals like *T-ED*, the IEEE Transactions on Circuits and Systems—I, IEEE Transactions on VLSI Systems, IEEE Transactions on Computer-Aided Design , and *TC*.