Analysis of Pattern-dependent Rapid Thermal Annealing Effects on SRAM Design

Vidya A. Chhabria[1] and Sachin S. Sapatnekar[2]
[1]Arizona State University; [2]University of Minnesota

***Abstract*—Rapid thermal annealing (RTA) is an important step in semiconductor manufacturing. RTA-induced variability due to differences in die layout patterns can significantly contribute to transistor parameter variations, resulting in degraded chip performance and yield. The die layout patterns that drive these variations are related to the distribution of the density of transistors (silicon) and shallow trench isolation (silicon dioxide) across the die, which result in emissivity variations that change the die surface temperature during annealing. While prior art has developed pattern-dependent simulators and provided mitigation techniques for digital design, it has failed to consider the impact of the temperature-dependent thermal conductivity of silicon on RTA effects and has not analyzed the effects on memory. This work develops a novel 3D transient pattern-dependent RTA simulation methodology that accounts for the dependence of the thermal conductivity of silicon on temperature. The simulator is used to both analyze the effects of RTA on memory performance and to propose mitigation strategies for a 7nm FinFET SRAM design. It is shown that RTA effects degrade read and write delays by 16% and 20% and read static noise margin (SNM) by 15%, and the applied mitigation strategies can compensate for these degradations at the cost of a 16% increase in area for a 7.5% tolerance in SNM margin.**

I. Introduction

Variations in the semiconductor manufacturing process cause significant degradation in yield and performance. One source of variation is the rapid thermal annealing (RTA) step in semiconductor manufacturing, which is widely adopted in ultra-shallow junction technologies to activate dopants after implantation. The process determines the dopant activation and lateral source/drain diffusion, which in turn dictates the threshold voltage, V_{th}. Since V_{th} directly affects circuit performance, it is crucial to ensure that the V_{th} of every transistor only shows small variations from the nominal V_{th}. While some level of random dopant fluctuation is inevitable and can lead to V_{th} drift, it is important to limit systematic variations induced by RTA by aiming for uniform temperatures across the entire die during the annealing process. However, traditional design flows do not directly address this source of variation. Instead, they rely on using design corners and adding heuristically-defined margins, which may be optimistic, leading to reduced yield or pessimistic, leading to lower performance per watt.

RTA is typically performed in a chamber where an energy source such as a lamp or laser emits energy to the surface of the wafer for a short duration, as shown in Fig. 1 (left bottom). By bringing the wafer to a high temperature, the process performs dopant activation while limiting unwanted dopant diffusion. Since radiation is the primary mechanism of heat transfer in the chamber, the temperature on the wafer depends

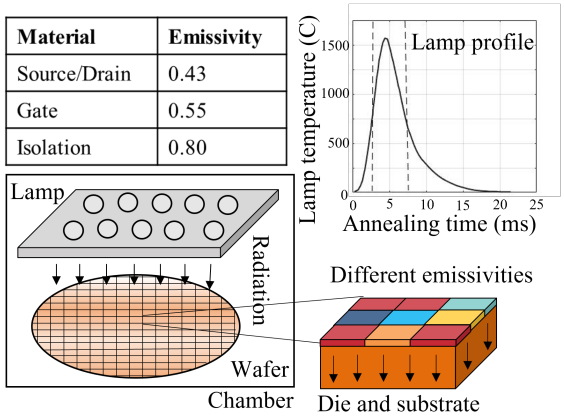

Fig. 1. RTA process highlighting wafer heating through radiation, the temperature profile of the lamp with millisecond anneal times and a 5ms pulse, and intra-die emissivity pattern variations.

on the emissivity of the surface. Therefore, different regions of the die absorb different amounts of heat based on their emissivity. Fig. 1(right bottom) shows a die with nine regions, each having different emissivities, resulting in different die anneal temperatures, and consequently, different V_{th} across different parts of the chip. Since the emissivity of each region depends on layout choices during design, it is important to account for these variations prior to manufacturing, to minimize RTA-induced systematic V_{th} variations.

Prior art has addressed this problem during design by analyzing and minimizing RTA-induced variations through the development of (i) pattern-dependent RTA simulators [1]–[3], (ii) compact predictive models for mapping anneal temperatures to transistor threshold voltages [1], (iii) physical design techniques to minimize emissivity gradients across the die [4], and (iv) optimal lamp profiles [5]. However, one of the shortfalls with existing commercial [6] and non-commercial simulators [1]–[3] is that they do not account for the dependence of the thermal conductivity of silicon on temperature. Given the high temperatures at which RTA is performed, this dependence is critical to account for as silicon changes from being a conductor at room temperature to being an insulator above $1000°$C. Further, the RTA-induced variation mitigation techniques that optimize the floorplan of a digital circuit [4] do not translate to memory design as the floorplan is set in stone. In this work, we develop a novel FDM-based RTA simulator that accounts for pattern-dependent effects and temperature-dependent thermal conductivity of silicon to estimate the die anneal temperature.

SRAMs are especially susceptible to RTA variation effects

as they have large variations in their layout patterns. Moreover, since SRAMs have stringent constraints on stability and performance, they have a low tolerance for variations in threshold voltages. An SRAM consists of dense arrays and sparse peripheral circuitry and is often surrounded by combinational control logic modules, all of which have different transistors densities, making the SRAM and its neighborhood particularly susceptible to RTA effects. In this work, we evaluate our simulator on a 7nm FinFET 16KB SRAM and propose RTA-induced variation mitigation techniques during design. We make the following contributions:

- We develop a 3D pattern-dependent transient RTA simulator that takes input GDS, material properties, and lamp properties to generate a thermal profile of the die.
- We demonstrate how the dependence of the thermal conductivity of silicon on temperature is critical for analyzing RTA effects as it increases the variation.
- We show that RTA-induced variations can degrade SNM by 17% and read and write latencies by 20%.
- We apply RTA-induced variation mitigation strategies, which add dummy SRAM 6T columns to prevent SNM degradation and peripheral circuitry sizing to compensate for the delay degradation. The compensation techniques come at the cost of a 16% increase in area for a 7.5% read SNM tolerance margin.

II. Preliminaries

A. RTA process description

RTA is a critical part of the semiconductor manufacturing process, performed after ion implantation, for the removal of the ion-implantation-induced damage and the activation of dopants. Unlike traditional annealing, RTA applies a much shorter pulse (e.g., lamp RTA [7]), to heat the silicon wafer to a much higher temperature. Fig. 1(right top) shows the profile of the lamp, which has a 5ms pulse and a maximum temperature of $1300-1800^{\circ}$C [1], [3], [8]. The basic mechanism for heat transfer during RTA is to rapidly transfer a large heat flux to the silicon die through radiation. The heat then spreads on the die by conduction. However, with the short annealing times (millisecond duration [8]), complete thermal equilibrium between conduction and radiation cannot be achieved, and the surface temperature is primarily determined by the ability of different regions of the die to absorb heat, i.e., the *emissivity* of the regions. The emissivity of the gate is usually higher than that of the source/drain region, while the isolation region has the highest emissivity due to the shallow trench isolation (STI) structure [9]. The difference in emissivity is highlighted in Fig. 1 (top left). As a result, different layout pattern densities lead to different annealing temperatures, which affects doping activation leading to differences in threshold voltages V_{th}, and consequently, performance. Fig. 1 shows the difference in emissivity, which leads to variations.

B. Compact model for RTA variations

Due to the differences in emissivity and the consequent differences in annealing temperature, a shift in threshold voltage and delay has been observed in test circuits [7]. Two primary independent mechanisms cause changes in threshold voltage due to RTA-based variations in anneal temperature: (i) dopant activation and (ii) lateral source/drain diffusion. These two effects together impact effective oxide thickness (EOT), threshold voltage (V_{th}), and effective junction depth L_{eff}. We use the compact model from [1], which relates RTA variations to changes in V_{th}, as summarized in Table I. The last row highlights the change in threshold voltage as a function of EOT and L_{eff}. The table shows the functional dependence of EOT on the doping activation, denoted by N_{act}. In the table, N_{max} refers to the maximum concentration of activated dopants; N_{min} is the minimum activated doping concentration, which refers to the activated doping concentration before the annealing; τ refers to the activation time constant, which is defined at the time that 50% dopants activated; t_{eff} is the effective annealing time such that the activation rate is equivalent to that of the simulated temperature profile. The values of these constants are typically available from RTA process parameters.

TABLE I
Compact model for V_{th} prediction under the RTA process [1].

Thermal Annealing
Dopant Activation
$N_{\text{act}}(T,t) = N_{\text{max}} + (N_{\text{min}} - N_{\text{max}})e^{-\frac{t_{\text{eff}}(T,t)}{\tau}}$
$t_{\text{eff}} = \int_0^t e^{\frac{E_a}{k(T^{-1}-T'(t)^{-1})}} dt$
S/D Lateral Diffusion (to define L_{eff})
$\Delta X_j(T,t) = a(T-T_0)^b(\sqrt{D(t-t_0)} - \sqrt{Dt_0})$
$L_{\text{eff}}(T,t) = L_{\text{eff}_0} - 2\Delta X_j(T,t)$
Device Parameters
$\text{EOT}(T,t) = T_{ox} + T_{poly} = T_{ox} + \frac{a}{N_{\text{act}}(T,t)}$
$V_{th}(T,t) = V_{\text{ref}} + (a + bL_{\text{eff}}(T,t))\text{EOT}(T,t)$

The second source of threshold voltage variation during the RTA process is junction depth, L_{eff}. The change in junction depth, ΔX_j, is modeled as a fitted polynomial as shown in the table, where T_0 is a reference temperature where the junction move is zero in a typical RTA process. The junction depth has a square-root dependence on the product of the diffusion coefficient D and annealing time t, where t_0 is the equivalent time before RTA.

C. SRAM Design

An SRAM and its peripheral circuitry are particularly susceptible to RTA effects due to large variations in layout density. The array has a higher transistor density when compared to the peripheral circuitry and therefore has a larger density of silicon which has a lower emissivity, as shown in Fig. 1 (left top) compared to the isolation regions.

Fig. 2(a) shows the top-level architecture of the SRAM with the locations of its blocks, and Fig. 2(b) shows the emissivity map generated GDS of the SRAM implemented in a FinFET 7nm technology. The emissivity of the row circuitry is higher when compared to the SRAM as it is sparser in the transistor density and denser in isolation regions, as shown in Fig. 2(c).

In this figure, the SRAM architecture consists of four arrays of 6T SRAM cells and peripheral circuitry, including decoders (row and columns), sense amplifiers, and one-shot circuitry

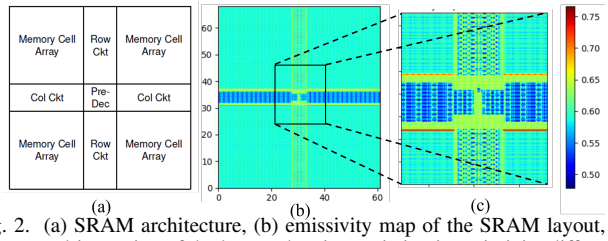

Fig. 2. (a) SRAM architecture, (b) emissivity map of the SRAM layout, and (c) zoomed-in version of the layout showing variation in emissivity differences between array and peripheral circuitry.

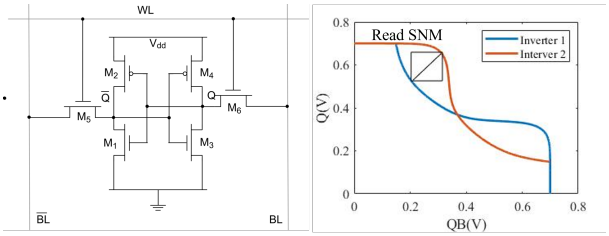

Fig. 3. (a) SRAM 6T cell and (b) butterfly curve for read SNM calculation.

for sense amplifier enable, precharge circuitry, and word line enable. The array consists of 6T SRAM cells made of two cross-coupled inverters to store bits and two access transistors for read and write, as shown in Fig. 3(a). The word line controls the access transistors, and the bit lines act as inputs during write and outputs during read operations.

The SRAM is designed to meet stability and performance constraints. The former relates to a static noise margin (SNM) constraint for every 6T SRAM cell, and the latter to read/write latency constraints. SNM is the least noise voltage needed to change the cell state. It is typically estimated by plotting a butterfly curve by drawing and mirroring the inverter characteristics and then finding the maximum possible square between them. The length of the side of the square gives SNM. For example, Fig. 3(b) shows the butterfly curve [10] for the 7nm technology SRAM 6T cell in Fig. 3(a) designed to have a read SNM of 135mV.

The read SNM is sensitive to the change in threshold voltage as shown by the different curves in Fig. 3 for different threshold voltages of the six transistors. A simultaneous increase in both NMOS and PMOS transistor threshold voltages increases SNM as the inverter is less susceptible to static noise, while lower threshold voltages result in a decrease. Further, the read/write latency is directly affected by changes in threshold voltages, where higher threshold voltages increase the latency.

III. Identifying Pattern-Dependent Temperature Nonuniformities Based on RTA Simulation

In this section, we describe an approach for simulating the effects of emissivity pattern variations on the die on the temperature distribution on the die during RTA. Our pattern-dependent RTA simulator models the heat transfer from the lamp to the die, taking into account transient effects, including the impact of the change in thermal conductivity with temperature, an effect that has not previously been captured by prior RTA simulation works [1]–[3]. It generates thermal profiles which account for pattern-dependent effects, temperature-dependent thermal conductivity, and boundary conditions. The output of the simulator is the 3D spatial distribution of temperature across the die and substrate as a function of time. The following inputs are required to perform this simulation:

(1) Design GDS which specifies the location of every transistor and isolation region to generate emissivity patterns.

(2) Lamp profile including the maximum anneal temperature, lamp temperature rise and fall rates, and anneal pulse times.

(3) Material parameters includes the emissivity and thermal conductivity of silicon and silicon dioxide (isolation).

A. Heat transfer model for RTA process

As described in Section II, the RTA process takes place in a chamber through radiation and conduction mechanisms. The underlying heat transfer is governed by the second-order partial differential equation as shown below:

$$c_p\rho\frac{\partial T}{\partial t} = k\left(\frac{\partial^2 T}{\partial x^2}+\frac{\partial^2 T}{\partial y^2}+\frac{\partial^2 T}{\partial z^2}\right)+\frac{\sigma\epsilon}{h}\left(T_{\text{lamp}}^4 - T^4\right) \quad (1)$$

The left-hand side of (1) represents the time variation of temperature, T, on the die, c_p is the specific heat, and ρ is the density of the die. The right-hand side consists of two terms: the first is related to conduction (which multiplies the thermal conductivity k of the die by the second spatial derivative of T) [11], and the second is related to radiation (given by the Stefan-Boltzmann law) where σ is the Boltzmann constant, and h is the thickness of the die. The heat source from the lamp can be represented either in terms of the lamp temperature T_{lamp} or as the incident heat. ϵ represents the emissivity of the die, which is pattern-dependent, as shown in Fig. 1(top left).

Finite difference method (FDM): To solve the heat equation numerically, we apply the finite difference method (FDM) [11] approach to solve (1), discretizing the volume of the die as shown in Fig. 4 into small-sized elements. The discretized element at location (x, y, z) has its emissivity ($\epsilon_{x,y,z}$), thermal conductivity ($k_{x,y,z}$), and temperature ($T_{x,y,z}$). Each spatial derivative term in (1) is replaced by a finite difference term. For example, for the x direction:

$$\frac{\partial^2 T}{\partial x^2} \approx \frac{1}{\Delta x^2}(k_{x,x+\Delta x,y,z}(T_{x+\Delta x,y,z} - T_{x,y,z}) + k_{x,x-\Delta x,y,z}(T_{x-\Delta x,y,z} - T_{x,y,z})) \quad (2)$$

where Δx is the size of the discretized element in the x direction, and $k_{x,x\pm\Delta x,y,z}$, the thermal conductivity of the neighboring element is multiplied by the difference in temperature between the two neighboring elements.

As a result, the partial differential equation (1) loses all spatial derivative terms and becomes an initial value problem represented by a first-order differential equation, where $\frac{\partial T}{\partial t}$ is represented as an algebraic function of temperature with an initial condition. Therefore, the RTA simulator solves the initial value problem to obtain the die temperature during RTA.

Temperature-dependent thermal conductivity of silicon: One key contribution of our work is that our RTA simulator considers the non-linear variation of the thermal conductivity of silicon with temperature during the RTA process as shown in Fig. 5. Under the log scale, at high temperatures, the thermal

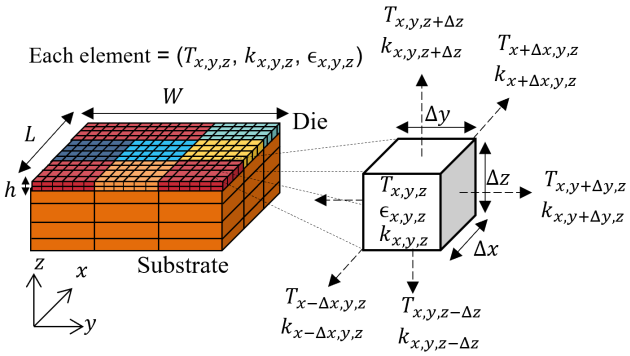

Fig. 4. FDM-based spatial derivative estimation.

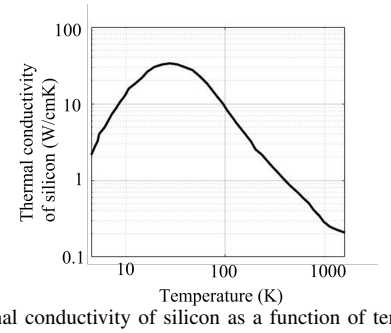

Fig. 5. Thermal conductivity of silicon as a function of temperature [12].

conductivity of silicon, which is an excellent heat conductor at room temperature, approaches that of an insulator [12].

Considering this non-linear dependency of temperature-dependent thermal conductivity is critical to accurately estimate the temperature of the die during the high-temperature anneals within the temperature. For example, Fig. 6(a) and (b) show the thermal profile of a testcase from [13] with and without considering the non-linear dependence of the thermal conductivity on temperature. When temperature-dependent thermal conductivity is considered, as in Fig. 6(a), radiation dominates conductance, and there is a larger variation in die temperature. Prior work [1]–[3] does not consider this dependence and hence incorrectly observes lower temperature variation with conduction playing a larger role. The larger temperature variation is crucial to capture as they cause significant RTA effects. We model this dependence through by curve-fitting of data from Fig. 5 to a ratio of polynomials, as shown by the following equation:

$$k(T) = \frac{p_1 T^2 + p_2 T + p_3}{q_1 T^2 + q_2 T + q_3} \tag{3}$$

where p_i and q_i are fitting coefficients and $k(T)$ is the thermal conductivity of silicon as a function of temperature T. We use this regression fit in (1) and (2) of our FDM model.

B. RTA simulation framework

To estimate the temperature profile on the die, our RTA simulation framework operates under the following assumptions: (i) Since RTA is applied at the wafer-level, for any analysis at the die-level, we can assume that the lamp is much larger when compared to the die, and radiation is incident perpendicular to its surface and uniformly distributed across the die as highlighted in Fig. 1.

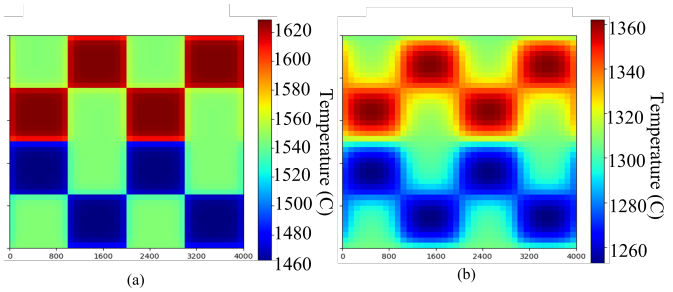

Fig. 6. Thermal profile of the testcase from [13] with a checkerboard pattern of emissivity (a) with and (b) without considering the dependence of the thermal conductivity of silicon on temperature. A lamp temperature of 1800°C and pule time of 5ms is used.

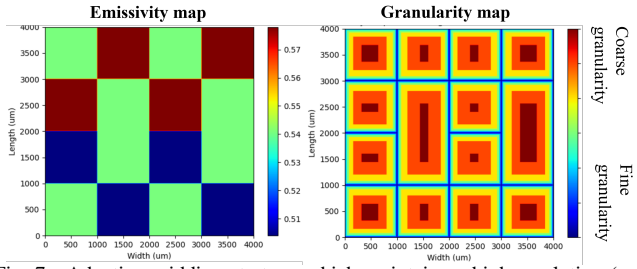

Fig. 7. Adaptive gridding strategy which maintains a high resolution (small element sizes) in the die regions with emissivity gradients. The testcase from [13] with four different element sizes.

(ii) At high temperatures, radiation dominates convection as the heat transferred by radiation is directly proportional to the fourth power of temperature, while the heat transferred by convection is proportional to the temperature. Therefore, we assume no heat transfer via convection.

Volumetric average to estimate $k_{x,y,z}$ and $\epsilon_{x,y,z}$ The die is discretized into small 3D elements, as shown in Fig. 4. The thermal conductivity $k_{x,y,z}$ and the emissivity $\epsilon_{x,y,z}$ of each element are obtained by taking the volumetric average of the thermal conductivity and emissivity of all the materials in the 3D element. The emissivity and thermal conductivity profiles of the die are obtained from the input design GDS and material properties of silicon and silicon dioxide. We use per-element thermal conductivity and emissivity to solve the FDM-based initial value problem in (1) using (2).

Boundary condition Based on the Fig. 1, each wafer has several dies patterned in a grid-like manner where each die, $D_{i,j}$, represents an integrated circuit (IC) at location (i, j) on the wafer. Therefore, ignoring the dies on the edge of the wafer, the temperature of the die, $D_{i,j}$ at $y = W$ is identical to the temperature of the die, $D_{i,j+1}$, at $y = 0$. Similarly, the temperature of die, $D_{i,j}$ at $x = L$ is the same as die, $D_{i+1,j}$ at $x = 0$. Similar to [4], we implement a circular convolution technique where we ensure the temperature at the $y = W$ edge of the die is the same as the temperature at the $y = 0$ edge of the die, and the temperature at the $x = W$ edge is the same as the $x = 0$ edge. Further, we model the path to thermal ground via the substrate of the wafer through a user-defined thermal conductance parameter.

Adaptive discretization Since the thermal analysis is performed using an FDM-based technique, there is a tradeoff between runtime and accuracy based on the element size,

where larger element sizes result in faster simulation times, but smaller element sizes result in accurate temperature estimates. Therefore, in our approach, we leverage an adaptive discretization strategy, where the smallest element size is an input to the simulator, and we estimate an initial value of $k_{x,y,z}$ of $\epsilon_{x,y,z}$ for each element. Our adaptive strategy uses progressively increasing filter sizes that estimate the gradient of the emissivity of the elements with its eight neighboring elements. If the gradient is zero, these are viable candidates to merge into a larger element. We perform this iteratively until no more elements can be merged. For example, for the emissivity map shown in Fig. 7(left), we obtain a granularity map shown in Fig. 7(right) where the blue regions indicate a high resolution (fine granularity, i.e., small element sizes), and the red regions indicate a low resolution (coarse granularity).

IV. Pattern-dependent RTA effects on SRAM

The RTA simulator is developed in Python3.7, which uses the initial value problem solver from the Scipy library. Our experiments are performed on Intel(R) Xeon(R) Gold 6132 CPU @ 2.60GHz with 480GB RAM. We evaluate the RTA simulator on a 7nm FinFET [14] 16KB SRAM in Fig. 2.

A. On die temperature variations during annealing

The die temperature depends on (i) layout patterns, (ii) the temperature-dependent conductivity of silicon, and (iii) the lamp profile (pulse time, temperature, etc.).

Impact of layout patterns As mentioned in Section II, the SRAM layout is extremely susceptible to RTA-induced variations due to differences in transistor density between the dense SRAM array and sparse peripheral circuitry and the consequent variation in emissivity across the die. Moreover, with the diversity of the blocks surrounding the memory e.g., combinational control logic, tap cell arrays, and filler cells, the variation in emissivity is larger, leading to significant RTA effects on the SRAM itself.

Fig. 8(a) shows the emissivity map extracted from the 16KB SRAM layout with combinational logic surrounding it on three sides and an array of filler and tap cells on the right. The regions around the SRAM without logic are occupied by filler cells and do not contain diffusion/active layers. Despite the design rule constraints (DRCs) requiring continuous horizontal fins and vertical gates across the whole die, regions without diffusion/active layers have an increased density of insulating silicon dioxide material which is nearly double the emissivity of conducting silicon (Refer Fig. 1(top left)). The low emissivity and thermal conductivity of silicon dioxide lead to high temperatures in those regions during RTA.

The thermal contours on the die are highlighted in Fig. 8(b), where the regions with a higher density of active (silicon) have lower die anneal temperatures. Fig. 8(c) shows the extracted temperature contours of the SRAM alone. These temperature contours are generated for a lamp profile specified in Fig. 1 with the 5ms anneal pulse time. For this testcase, our RTA simulator takes 40 minutes to generate in our setup.

Impact of temperature-dependent thermal conductivity The temperature contour plot in Fig. 8(b) and (c) is generated after considering the temperature-dependent thermal conductivity of silicon. In contrast, Fig. 8(d) is generated using a fixed thermal conductivity of silicon at room temperature. There is negligible RTA-induced variation in temperature ($\leq 0.6°$C) in Fig. 8(d) because, at high temperatures, the thermal conductivity of silicon approaches that of an insulator. Therefore, at high temperatures, radiation dominates conduction and increases variation due to pattern-dependent effects making it critical to account for in RTA simulations.

Impact of different lamp profiles Fig. 9(a) shows the maximum temperature on the die for a lamp anneal pulse of 5ms and 10ms. For the longer pulse time, the die temperature nearly reaches a steady state close to the lamp temperature ($1800°$C) and therefore has lesser temperature variation as shown in the temperature contour plot in Fig. 9(b). For the shorter pulse, the radiation component dominates conduction, and there is an increase in the temperature variation on the die, as shown in Fig. 8(c).

Scalability analysis We sweep the element size in our FDM simulation and plot the number of nodes versus runtimes and peak memory requirements, as shown in Fig. 10(a) and (b). With the increase in the number of nodes, the runtime and peak memory requirements increase non-linearly. However, even with small element sizes (50,000 nodes), our solver can estimate temperature within 7500s (about 2 hours). Our simulations are limited by peak memory and not runtime. For example, 50,000 nodes require the usage of 200GB RAM, and smaller element sizes go over the 480GB RAM limit of the machine on which we perform our experiments.

B. Threshold voltage variations

After generating the spatial thermal distribution on the die, we use the compact model described in Table I to predict the changes in threshold voltages for the die temperatures during annealing. For the lamp profile in Fig. 1 with a 5ms anneal pulse, we observe that the SRAM has a peak temperature of $1792°$C and a minimum temperature of $1747°$C at the 10ms time instance as shown in Fig. 8(c). Based on the choice of the reference temperature, T_0, i.e., the temperature on the die at which there is no change in threshold voltage, there are two possible worst-case scenarios:

Scenario-1: When T_0 is the minimum die anneal temperature, i.e., $T_0 = 1747°$C (Fig. 8(c)), the majority of the SRAM array and its peripheral have $\Delta V_{th} = 0$ (no change from the nominal value), as shown in Fig. 8(f). However, the SRAM cells and peripheral circuitry near the right edge of the array have a $\Delta V_{th} = -30$mV[1] which causes SNM degradation.

Scenario-2: When T_0 is the maximum on-die anneal temperature, i.e., $T_0 = 1792°$C (Fig. 8(c)), the majority of the SRAM array and its peripheral circuitry has a higher V_{th} than the reference. Fig. 8(e) shows a $\Delta V_{th} = 30$mV leading to an increase in latency. These changes in ΔV_{th} correspond to a 10% change in nominal V_{th}, which is a significant variation to cause performance and yield issues.

[1]For NMOS transistor, the threshold voltage is lower than what it is designed for, and for PMOS transistor it is higher.

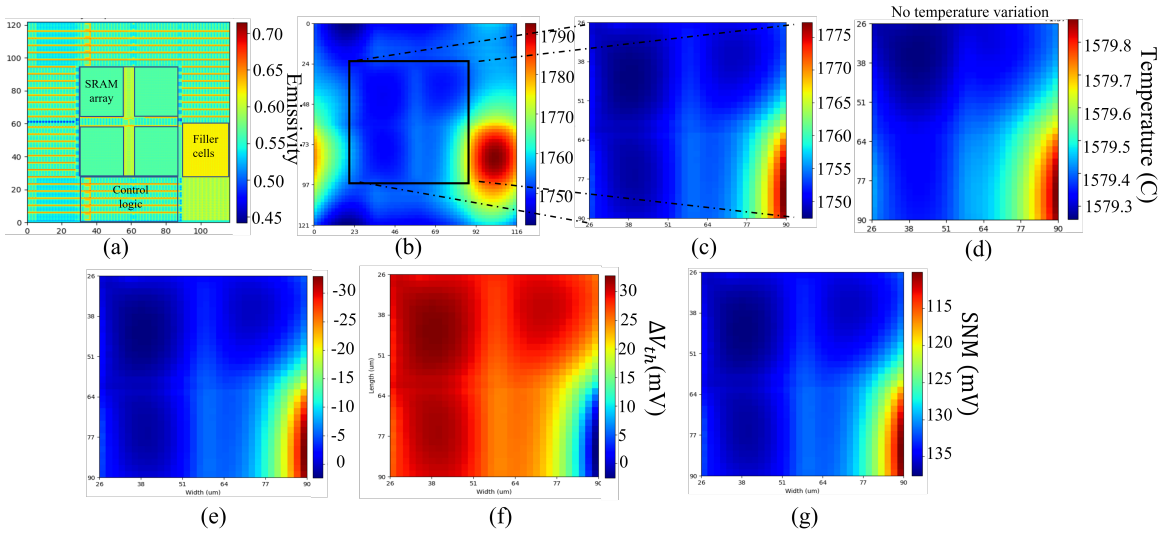

Fig. 8. (a) Emissivity map of the SRAM and its surrounding blocks. Temperature contour plot during annealing of: (b) the entire chip and (c) the SRAM block alone when the temperature-dependent thermal conductivity of silicon is considered, and (d) the SRAM block alone when the thermal conductivity of silicon at room temperature is used. The shift in threshold voltage in (e) Scenario-1 and (f) Scenario-2 due to different die temperatures during annealing. (f) Variation in SNM due to different die temperatures during annealing in Scenario-1.

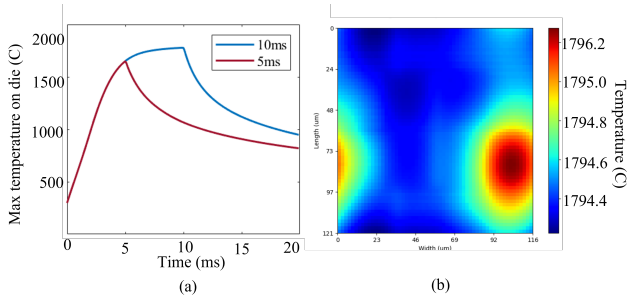

Fig. 9. (a) Maximum temperature on the die for two different pulse times for a 1800°C lamp and (b) temperature contours for the 10ms anneal pulse at time instance with the maximum temperature.

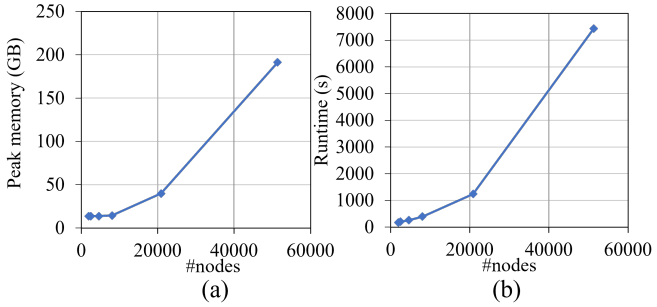

Fig. 10. Scalability analysis: (a) runtimes and (b) peak memory as a function of the number of nodes in the FDM simulation.

C. Impact on read SNM

The SRAM 6T cell in Fig. 3 is designed to meet the read SNM constraint of 135mV at all corners for the FinFET 7nm technology node [14]. However, when RTA effects are taken into the picture, the variation in the V_{th} of the transistor affects the read SNM. An increase in V_{th} of all transistors in the SRAM 6T cell increases SNM as it now requires a larger static noise to flip the state of the cell.

Unlike aging-related V_{th} degradations, which impact individual transistors differently in the 6T cell [15], RTA-induced V_{th} variations impact all transistors equally as the temperature within the SRAM cell is virtually the same. This is due to low local spatial temperature variations caused by the averaging of emissivities that filter out small local variations. Therefore, unlike aging-related V_{th} degradation, which asymmetrically impacts the 6T cell decreasing SNM, RTA-induced V_{th} increases improve the read SNM.

For Scenario-1, the array of SRAM cells at the right edge of the right-bottom quadrant have an SNM of 115mV, which is lower than the 135mV specification, as shown by the red regions in Fig. 8(g). However, For Scenario-2, the worst-case SNM is still larger than the constraint of 135mV, ensuring the stability of the cell.

D. Impact on read and write latency

We design and simulate the critical path of the 16KB SRAM such that it meets its SNM (135mV) and latency constraints (400ps read and write latency) at the slowest corner. This simulation provides us with the baseline implementation with read and write latency values as shown in the first column of Table II. The table lists delays of the predecoder, i.e., CLK to predecoder and delay of the row decoder and predecoder together, i.e., CLK to word line (WL). It also lists the precharge circuitry delays for read and write operations.

With RTA-induced variation, the change in threshold voltages impacts the transistors in both the peripheral and SRAM array transistors. To estimate the impact on read and write latency, we annotate the estimated ΔV_{th} back into the SPICE netlist and simulate it. We perform this simulation only for Scenario-2 (Fig. 8(f)) as the increase in threshold voltages will degrade the delays. The change in read and write latencies after accounting for RTA-induced variations are listed in Table II for the same slow corner at which it was designed. The variations increase the threshold voltages of all transistors on the critical path by $\approx$ 10%, which results in the read and write latency increasing by 16% and 20%, violating the constraints for the read operation.

TABLE II
READ AND WRITE LATENCY WITH AND WITHOUT CONSIDERING RTA-INDUCED ΔV_{th} VARIATIONS AND AFTER DESIGNING PERIPHERAL CIRCUITS TO MITIGATE (MIT.) RTA VARIATIONS.

	Without RTA ΔV_{th}	With RTA ΔV_{th}	With RTA ΔV_{th} mit.
CLK to predecoder	38.95 ps	46.77 ps	41.83 ps
CLK to word line	96.43 ps	117.64 ps	103.54 ps
Read latency	371.28 ps	430.68 ps	396.17 ps
Write latency	315.62 ps	379.11 ps	352.18 ps
Precharge delay read	341.83 ps	403.35 ps	359.56 ps
Precharge delay write	357.17 ps	414.24 ps	388.79 ps

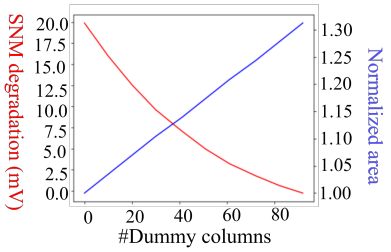

Fig. 11. Area and SNM tradeoff due to insertion of dummy columns.

V. MITIGATION OF RTA EFFECTS DURING SRAM DESIGN

We apply two RTA-induced variation mitigation strategies for SRAM design. First, to compensate for the degradation in read SNM in Scenario-1, we propose the addition of dummy SRAM columns between the array and the surrounding blocks. Second, to compensate for the increase in latency in Scenario-2, we increase the sizes of transistors in the peripheral circuitry to improve drive strengths and reduce delays.

A. Addition of dummy SRAM 6T columns

Based on the estimated temperature and predicted SNM gradients across the SRAM, we add the dummy SRAM columns in critical locations. These dummy columns are unused sacrificial 6T cells, which are placed in the regions that have the most degraded SNMs, allowing the non-dummy 6T cells to operate within the stability constraints.

For the testcase in Fig 8(a), and the generated SNM contours in Fig. 8(g), we find that the SRAM cells in the lower right quadrant of the SRAM array to have degraded SNMs. Particularly, the 60 columns between $x = 70$ to $x = 90$ locations have lower than 135mV read SNM. Ideally, to have all SRAM 6T cells meet the constraint for this testcase, we need to add 60 additional columns. However, this would increase the area of the SRAM by $\approx 1.2\times$.

Based on the allowable tolerance margins for read SNM, the number of dummy columns to be inserted can be selected to balance the tradeoff between area and SNM. Fig. 11 plots the tradeoff between degradation in SNM, the area of the array (normalized to the area without dummy cells), and the number of dummy columns inserted. For example, for an allowable limit of 10mV degradation in SNM, we can add 27 dummy columns each of width 0.216μm, which increases the area by $1.16\times$. There is no impact on delays due to the additional columns as they do not lie on the critical path.

B. RTA-aware peripheral circuitry design

To compensate for the increase in read and write latencies, we redesign the peripheral circuitry to meet the latency constraints after considering the ΔV_{th} variations due to RTA. We change the timing of the one-shot circuitry by changing the number of inverter stages and the transistor sizes of the other components in the peripheral circuitry. In particular, we increase the size of the inverters in the predecoder and row decoder from $4\times$ to $6\times$ sized gates and increase the number of fins in the transistors of the read and write multiplexers by $1.2\times$ to reduce the latency in the read and write path. In addition, we increase the number of fins in the precharge circuit by $1.33\times$. The new delays are listed in the third column of Table II and show that the new latencies meet the design constraints. The simulator described in Section III allows for the exploration of transistor sizes which can mitigate RTA latency effects by providing designers an estimate of how much delay to recover.

These changes do not contribute to an increase in the area of the SRAM block as they are made to the peripheral circuitry, which is underutilized and any increase in the number of fins is absorbed. Further, the overall area of the SRAM block is dominated by the four SRAM arrays, as shown in Fig. 2(a).

VI. CONCLUSION

We develop a pattern-dependent transient RTA simulator that considers the temperature-dependent thermal conductivity of silicon. The simulator is used to analyze the RTA effects on a 16KB SRAM and to suggest variation mitigation strategies.

Acknowledgments: This work is supported in part by SK Hynix. We would like to thank Prof. Chris Kim and Meghna Madhusudan for helping with testcases and simulations.

REFERENCES

[1] Y. Ye, *et al.*, "Variability analysis under layout pattern-dependent rapid-thermal annealing process," in *Proc. DAC*, pp. 551–556, 2009.
[2] M. Rabus, *et al.*, "Rapid thermal processing of silicon wafers with emissivity patterns," *Journal of Electronic Materials*, vol. 35, pp. 877–891, 12 2006.
[3] E. H. Granneman, *et al.*, "3D pattern effects in RTA radiative vs. conductive heating," *ECS Transactions*, vol. 3, p. 85, oct 2006.
[4] Y. Wei, *et al.*, "Physical design techniques for optimizing RTA-induced variations," in *Proc. ASP-DAC*, pp. 745–750, 2010.
[5] R. Gunawan, *et al.*, "Optimal control of rapid thermal annealing in a semiconductor process," *Journal of Process Control*, vol. 14, no. 4, pp. 423–430, 2004.
[6] COMSOL, "Rapid Thermal Annealing." https://www.comsol.com/model/rapid-thermal-annealing-504.
[7] T. Gebel, *et al.*, "Millisecond annealing with flashlamps: Tool and process challenges," in *Proc. RTP*, pp. 47–55, 2006.
[8] P. Timans, *et al.*, "Millisecond annealing: Past, present and future," *MRS Proceedings*, vol. 912, pp. 0912C01–01, 2006.
[9] B. Walsh, *et al.*, "RTA-driven intra-die variations in stage delay, and parametric sensitivities for 65nm technology," in *Proc. VLSIT*, pp. 170–171, 2006.
[10] B. Calhoun and A. Chandrakasan, "Static noise margin variation for sub-threshold SRAM in 65-nm CMOS," *IEEE Journal of Solid-State Circuits*, vol. 41, no. 7, pp. 1673–1679, 2006.
[11] Y. Zhan, *et al.*, "Thermally-aware design," *Found. Trends Electron. Des. Autom.*, vol. 2, pp. 255–370, Mar. 2008.
[12] E. Bilotti, *et al.*, "Organic thermoelectric composites materials," in *Comprehensive Composite Materials II*, pp. 408–430, Oxford: Elsevier, 2018.
[13] K. L. Knutson, *et al.*, "Physical modeling of layout-dependent transistor performance," *ECS Transactions*, vol. 13, p. 63, oct 2008.
[14] L. T. Clark, *et al.*, "ASAP7: A 7-nm FinFET predictive process design kit," *Microelectronics Journal*, vol. 53, pp. 105–115, 2016.
[15] S. Kumar, *et al.*, "Impact of NBTI on SRAM read stability and design for reliability," in *Proc. ISQED*, pp. 6 pp.–218, 2006.