Analyzing electrical effects of RTA-driven local anneal temperature variation

Abstract – Suppresing device leakage while maximizing drive current is the prime focus of semiconductor industry. Rapid Thermal Annealing (RTA) drives process development on this front by enabling fabrication steps such as shallow juction formation that require a low thermal budget. However, decrease in junction anneal time for more aggresive device scaling has reduced the characteristic thermal length to dimensions less than the typical die size. Also, the amount of heat transferred, and hence the local anneal temperature, is affected by the layout pattern dependence of optical properties in a region. This variation in local anneal temperature causes a variation in performance and leakage across the chip by affecting the threshold voltage (V_{th}) and extrinsic transistor resistance (R_{ext}). In this work, we propose a new local anneal temperature variation aware analysis framework which incorporates the effect of RTA induced temperature variation into timing and leakage analysis. We solve for chip level anneal temperature distribution, and employ TCAD based device level models for drive current (Ion) and leakage current (Ioff) dependence on anneal temperature variation, to capture the variation in device performance and leakage based on its position in the layout. Experimental results based on a 45nm experimental test chip show anneal temperature variation of up to ~10.5^{o}C, which results in ~6.8% variation in device performance and ~2.45X variation in device leakage across the chip. The corresponding variation in inverter delay was found to be ~7.3%. The temperature variation for a 65nm test chip was found to be ~8.65^{o}C.

1. INTRODUCTION

Semiconductor industry faces significant challenges as it strives to extend Moore's law through aggressive process scaling. The most important challenge lies in maximizing the device on-current, while suppressing the leakage. Such a progress is driven by advances in the engineering of ultra-thin gate insulators, high-mobility channels, ultra-shallow junctions and low-resistance contacts. RTP (Rapid Thermal Processing) is a key process step in providing the essential capabilities for both process and material development on this front [1]. Figure 1 illustrates the important role of RTP in an advanced fabrication process.

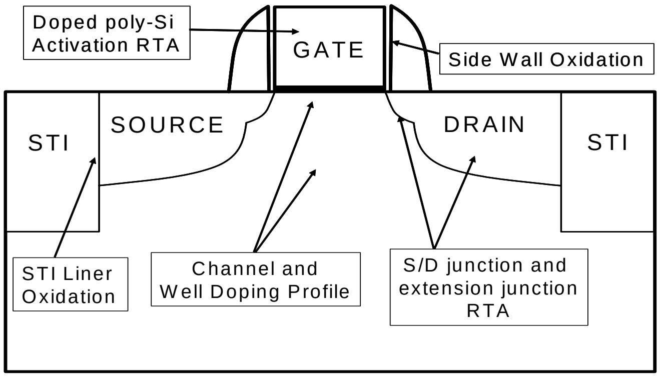

Figure 1. Role of RTP in advanced fabrication process.

RTP is employed in fabrication steps that require the wafer to be heated and cooled quickly within a low thermal budget (a small value of temperature time product) [2]. For example, the shallow p+-n junctions are difficult to fabricate due to high Boron diffusivity and formation of Boron channeling tail. Rapid Thermal Anneal (RTA) has been successfully used to address this problem [3, 4]. RTA typically involves spike anneal, where the wafer is ramped to a high temperature and then allowed to cool immediately [5]. Spike anneal allows the use of high temperature for higher dopant activation, and ion implantation damage annealing, but restricts the dopant diffusion by minimizing effective anneal time (low thermal budget).

However, decrease in the anneal time to achieve shallower junctions for more aggressively scaled transistors, has resulted in decreasing the characteristic thermal length (the length over which thermal equilibrium can be reached for a given time) to dimensions less than the typical die size [6]. In addition, since radiation is the primary source of heat transfer, layout pattern dependence of optical properties (emissivity, reflectivity, etc.) also affects the amount of heat absorbed and hence the local anneal temperature of a region in the layout [7]. This causes variation in the local anneal temperature across the chip, which in turn affects the transistor performance and leakage [8].

Higher local anneal temperature drives the junctions both longitudinally and vertically, and causes a higher activation of dopants. This results in lower threshold voltage (V_{th}) by a combination of short channel effects and compensation of halo doping. Also, higher dopant activation and increased gate overlap of source and drain together result in lower extrinsic transistor resistance (R_{ext}). This correlation in the across-chip variation of V_{th} and R_{ext} results in a pronounced effect on the drive current and leakage. Since the characteristic thermal length is quite large, entire circuit blocks may be systematically faster or slower depending on their position in the layout. Thus, not considering the RTA induced variations might result in significant misinterpretation of circuit timing. Experiments in [9] showed that ring oscillator frequency can vary by as much as ~20% based on the position in the die due to local anneal temperature variation.

There has been work in the past focusing on obtaining rigorous solution to local anneal temperature variation, and analyzing its effect on rapid thermal processes such as oxidation [10]. But the analysis in these works is very involved and has not been extended to a framework which can be used efficiently for any given layout. To the best of our knowledge, no work in the past has addressed the problem of obtaining quick and efficient solution to anneal temperature distribution, and use it for RTA aware

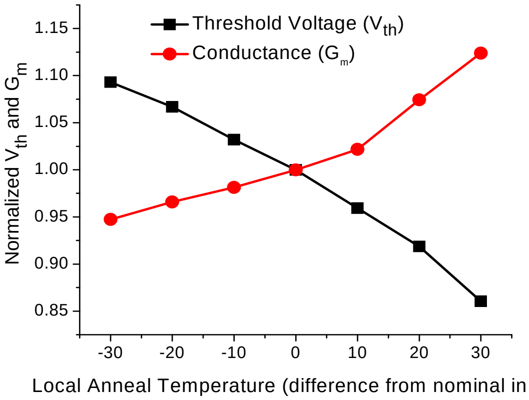

Figure 2. PMOS V_{th} and G_m variation with anneal temperature.

timing analysis. In this paper, we propose a new RTA aware timing framework which embodies transistor level models for anneal temperature sensitivity, to incorporate RTA induced temperature variation into traditional timing/leakage analysis. This is achieved by modeling the dependence of drive current (Ion), and leakage current (Ioff) on local anneal temperature, using device level TCAD simulations. Next, we mesh the wafer into rectangular grid, and solve for local anneal temperature using finite difference method. This involves discretization of the second spatial derivative of temperature as a finite difference in rectangular co-ordinates. Once the local anneal temperatures are known, we use multipliers for Ion and Ioff to enable accurate timing and leakage analysis for a device (based on its position in the wafer). Experimental results show that ignoring the intra-die variation due to RTA can lead to errors of up to 6.8% in device drive current (~7.3% in inverter delay), and ~2.45X in device leakage, for a 45nm test chip. The temperature variation for a 65nm experimental test chip was found to be ~8.65^{o}C.

The rest of the paper is organized as follows. Motivation and background for this work is discussed in Section 2. Section 3 presents the methodology and simulation flow for RTA aware timing analysis. Experimental results are discussed in Section 4, and Section 5 concludes the paper.

2. Motivation and Device Level Analysis

As discussed in Section 1, local anneal temperature varies with position on a die. This variation in turn affects device properties such as vertical and longitudinal position of the S/D junctions, dopant activation, and compensation of the halo doping. Higher local anneal temperature results in lower V_{th} and R_{ext} due to a combination of these effects, and hence faster devices. Prior works have reported the intra-die variation of R_{ext} and V_{th} to be highly correlated, and this makes the strength of this variation particularly strong.

Figure 2 illustrates the temperature dependence of threshold voltage (V_{th}) and conductance (G_m) for a 45nm PMOS device. As expected, V_{th} decreases, while G_m increases with increase in local anneal temperature. Figure 3 shows the effect of temperature variation on gate to source capacitance (C_{gs}) and drain induced barrier lowering (DIBL). Both C_{gs} and DIBL increase with increase in temperature as gate/drain and gate/source overlap increases upon increasing local anneal temperature due to increased dopant diffusion. In order to examine the effect on performance and leakage, drive current (Ion) and leakage current (Ioff) are also plotted as a function of temperature in Figure 4. In all of these plots, 0 denotes nominal anneal temperature, and the other temperature values denote the variation from the nominal value. All the other values plotted as a function of temperature, are normalized to the corresponding values at nominal anneal temperature.

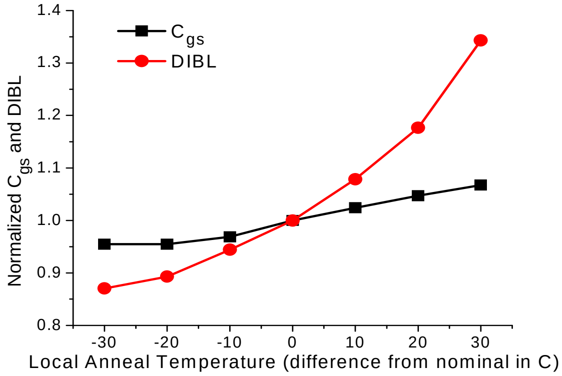

Figure 3. PMOS C_{gs} and DIBL variation with anneal temperature.

The plots clearly show significant dependence of performance and leakage on local anneal temperature, due to the strong dependence of V_{th}, G_m, C_{gs}, and DIBL on it. An increase as small as 20^{o}C over the nominal anneal temperature, results in ~11.8% change in Ion, and ~4X change in Ioff. Decreasing the anneal temperature by 20^{o}C results in ~7.4% and ~2.6X decrease in Ion and Ioff, respectively. Previous works [10] have reported anneal temperature differences of up to 50^{o}C, between the highest and lowest anneal temperature on the wafer, for regular patterns. Such a difference can cause significant variation in performance and leakage which must be accounted for while analyzing performance and leakage. The plots also suggest that anneal temperature dependence of Ion and Ioff can be modeled as polynomial functions of temperature. The next section develops these models, discusses the methodology for calculating local anneal temperature, and the simulation flow for the RTA aware timing analysis.

3. Methodology

There are two major components of RTA aware simulation framework: models for the effect of local anneal temperature on leakage and drive currents, and methodology to solve for local

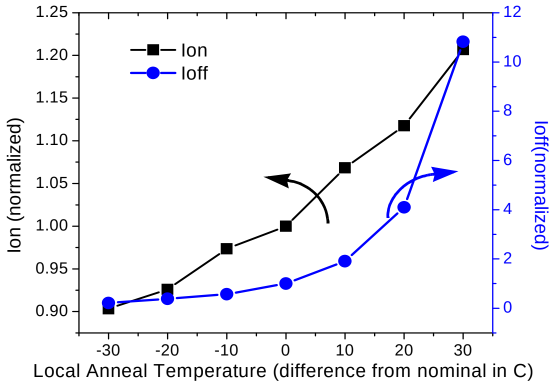

Figure 4. PMOS Ion and Ioff variation with anneal temperature.

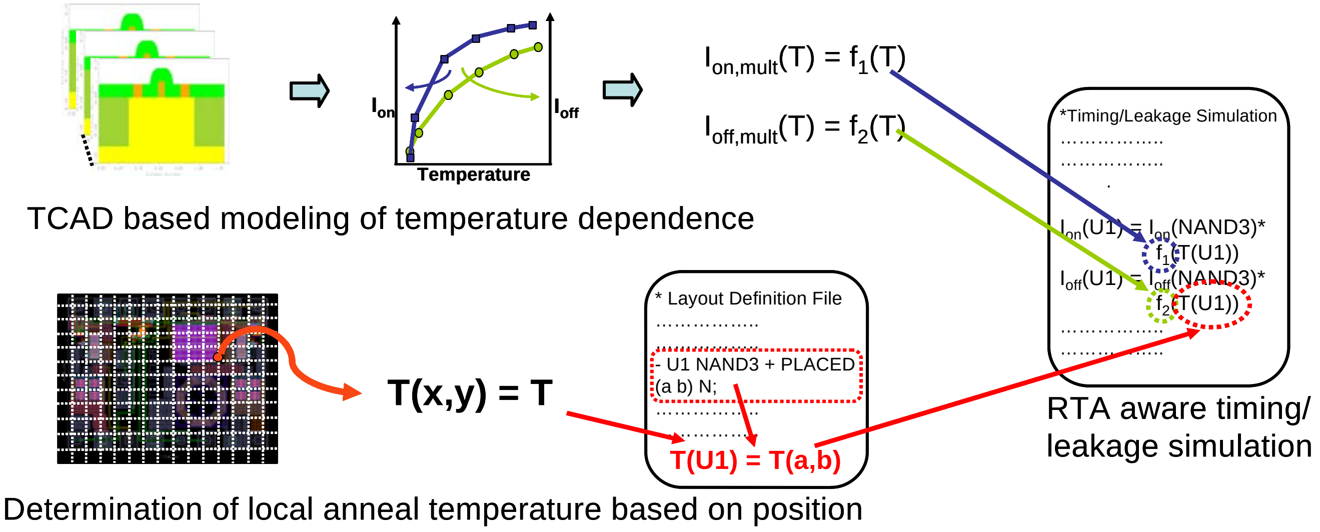
Ion
Ioff
Temperature
$I_{on,mult}(T) = f_1(T)$
$I_{off,mult}(T) = f_2(T)$
*Timing/Leakage Simulation
……………..
……………..
.
$I_{on}(U1) = I_{on}(NAND3)* f_1(T(U1))$
$I_{off}(U1) = I_{off}(NAND3)* f_2(T(U1))$
……………..
……………..
TCAD based modeling of temperature dependence
T(x,y) = T
* Layout Definition File
……………..
- U1 NAND3 + PLACED (a b) N;
……………..
……………..
T(U1) = T(a,b)
RTA aware timing/ leakage simulation
Determination of local anneal temperature based on position

Figure 5. RTA aware performance/leakage analysis flow.

anneal temperature. Figure 5 illustrates this simulation flow. Device level TCAD simulations are performed using TSUPREM4 in order to model the dependence of leakage current and drive current on local anneal temperature, in terms of on and off current multipliers. The layout is meshed into rectangular grid and finite difference method is used to solve for temperature as a function of position on the wafer. This temperature distribution is used to determine local anneal temperature for a gate using the layout definition file to determine its position. Once the local anneal temperature is known, the Ion and Ioff models can be used to modify the values for leakage and delay (read from the characterized library) at the time of final simulation. In this work, we assume that gate delay is inversely proportional to Ion, and hence we scale the gate delay by the inverse of the Ion multiplier. The remainder of this section describes in detail the development of these models, the method to solve for local anneal temperature, and concludes by discussing the RTA aware simulation flow.

3.1. Modeling performance and leakage variation with local anneal temperature variation

As discussed earlier, variation in local anneal temperature affects V_{th} and R_{ext} by a combination of changes in dopant activation, effective channel length, halo doping and gate overlap of source and drain. There can be two approaches to take these effects into account during timing and leakage analysis:

- Model the effects on basic device properties (V_{th} and R_{ext} or doping profile, effective channel length, and halo doping), use these models to modify the HSPICE model files, and characterize standard cell library at different anneal temperatures. For a given temperature interpolate between the known values from library files characterized at certain fixed values of local anneal temperature.
- Model the effects in terms of Ion and Ioff multipliers as a function of local anneal temperature. Characterize the standard cell library once for nominal anneal temperature, and superimpose these multipliers at the time of timing/leakage analysis when reading in the characterized values for a given standard cell.

Since the focus of this work is to perform accurate delay and leakage simulation, the second approach provides more accurate results than interpolation. Also, the second approach is easier to implement and has a lower characterization cost as compared to the first approach. So, we have used the second approach to construct the simulation framework in this work.

We use TCAD (TSUPREM4 [11] for process simulation, and MEDICI [12] for device simulation) for device level simulation and I-V characteristics of the devices are matched to the 45nm

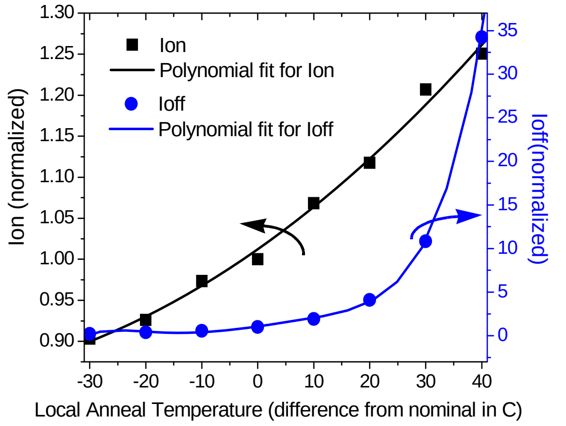

Figure 6. PMOS models for Ion and Ioff variation with temperature.

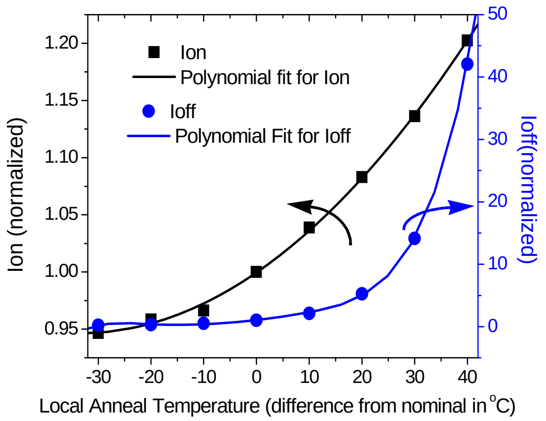

Figure 7. NMOS models for Ion and Ioff variation with temperature.

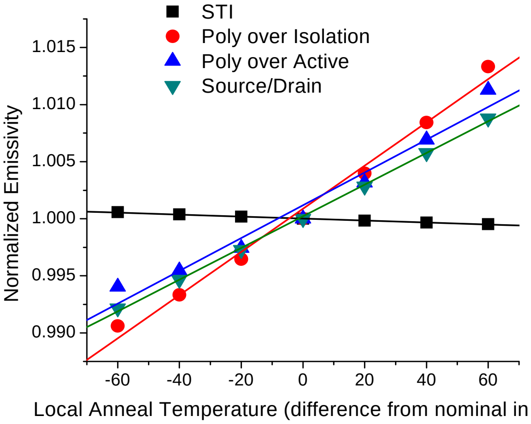

Figure 8. Model for emissivity variation with anneal temperature.

technology used for this work. Figure 6 shows normalized values of Ion and Ioff (normalized to the values at nominal anneal temperature) plotted against the anneal temperature for an NMOS device, and Figure 7 shows the corresponding plot for a PMOS device. Also shown are our polynomial models for Ion and Ioff as a function of temperature. Since these models are fitted to the normalized values, they represent multipliers by which nominal Ion and Ioff should be multiplied to obtain their correct values for a given temperature. The exact relationship between leakage and drive currents as anneal temperature is varied is a complex non-linear function. However, we see that they can be modeled with good accuracy using polynomial functions of the anneal temperature. We used MATLAB for accurate curve fitting, and observed that a quadratic polynomial can be used to model the Ion dependence with good accuracy, while a polynomial of degree 5 was needed to model the off current dependence, for both NMOS and PMOS. It is evident from the figures that polynomials predict the Ion and Ioff dependence accurately.

3.2. Chip level anneal temperature variation analysis

Solving for local anneal temperature involves setting up differential equations describing the heat flow, and then discretizing the chip area into rectangular grid and approximating the partial derivatives. Since radiation is the dominant mechanism of heat transfer, we assume that conduction and convection have negligible contribution. We also assume that the temperature variation across the thickness of the chip is negligible compared to the variations in the plane of the wafer. This assumption is based on the fact that characteristic time duration of RTA processes (in the order of a few seconds) is large enough to allow the temperature distribution across the thickness of the wafer to reach a study state. These assumptions are valid, and have been used in the past to make accurate predictions for RTP processes. This allows us to solve for the temperature in the plane of the wafer (x-y plane), and in particular we can assume the partial derivative with respect to the wafer thickness (z) to be zero.

In the steady state, we can write the heat balance equation as Poisson's equation [10]

$$\ddot{d}k\nabla^2 T(x,y) = -(P_{ABS}(x,y) - P_{EMI}(x,y)) \qquad (1)$$

where *d* is the wafer thickness, *k* is the in plane thermal conductivity, *T(x,y)* is the temperature distribution in the wafer plane, P_{ABS} is the radiative power absorbed per unit area, and P_{EMI} is

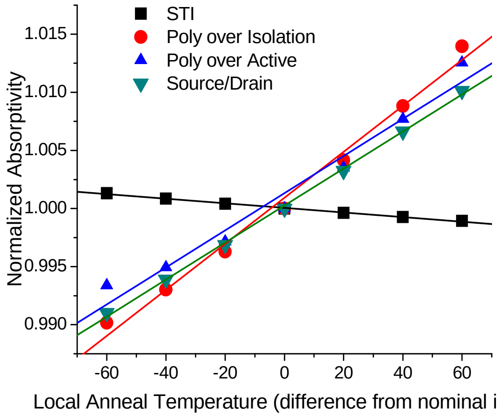

Figure 9. Models for absorptivity variation with anneal temperature.

the radiative power per unit area emitted by the wafer. The emitted and absorbed power perms vary with position on the wafer due to the layout pattern and hence position dependence of optical properties (emissivity, absorptivity, etc.), and can be expressed as

$$P_{ABS}(x,y) = \alpha(x,y,T)P_{incident} \qquad (2)$$

$$P_{EMI}(x,y) = \varepsilon(x,y,T)\sigma T^4(x,y) \qquad (3)$$

where $P_{incident}$ is the heater power per unit area incident on the wafer, σ is the Stefan-Boltzmann constant, $\alpha(x,y,T)$ is the position and temperature dependent effective total absorptivity, and $\varepsilon(x,y,T)$ is the effective wafer emissivity. The effective emissivity and absorptivity depend on optical properties of the layer structure, and upon the temperature of the region. They also depend on the wavelength of the radiations incident on the wafer, which in turn depends on the heater temperature and material (known for a given fabrication process).

The steady-state heat balance equation (1) can be discretized by writing the second derivative of temperature as a finite difference term. We use a grid based approach, where we discretize the wafer surface into a rectangular grid structure with lengths Δx, and Δy in the x and y direction respectively. This gives us one discritized node equation for each node on the grid. The spatial derivative of temperature in equation (1) can be written as

$$\nabla^2 T(x,y) = \frac{\partial^2}{\partial x}T(x,y) + \frac{\partial^2}{\partial y}T(x,y) \qquad (4)$$

and we can discritize the individual second spatial derivative terms. Let $T_{a,b}$ represent the anneal temperature at a node with co-ordinates (a,b). Now, in the x direction, we can write

$$\frac{\partial^2 T_{a,b}}{\partial x} \approx \frac{\frac{T_{a-1,b} - T_{a,b}}{\Delta x} - \frac{T_{a,b} - T_{a+1,b}}{\Delta x}}{\Delta x} \qquad (5)$$

$$\frac{\partial^2 T_{a,b}}{\partial x} \approx \frac{T_{a-1,b} - 2T_{a,b} + T_{a+1,b}}{(\Delta x)^2} \qquad (6)$$

Discretizing the heat balance equation yields the following system of non-linear equations (one equation for each node on the grid):

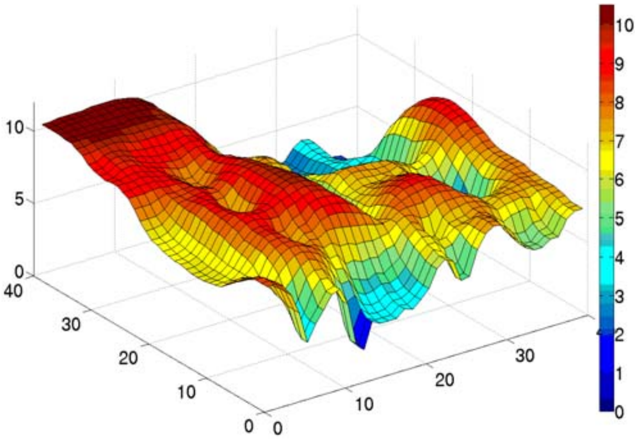

Figure 10. Local anneal temperature distribution for the 45nm chip.

$$\frac{T_{a-1,b} - 2T_{a,b} + T_{a+1,b}}{(\Delta x)^2} + \frac{T_{a,b-1} - 2T_{a,b} + T_{a,b+1}}{(\Delta y)^2} = \\ -(\alpha_{a,b}(T_{a,b})P_{incident} - \varepsilon_{a,b}(T_{a,b})\sigma T^4{}_{a,b}) \quad (7)$$

where $\alpha_{a,b}(T_{a,b})$, and $\varepsilon_{a,b}(T_{a,b})$ are the position dependent functions of temperature, describing the average behavior of absorptivity and emissivity in a rectangle with sides Δx, and Δy, centered at the node point. The optical properties at a given point depend on the layer structure of the wafer. There are five different kinds of layer structures possible at the time of RTA: N+ source/drain, P+ source/drain, polysilicon over isolation, polysilicon over transistor, and shallow trench isolation (STI). We use a tool called Rad-Pro [13] to calculate the temperature dependent functions of emissivity and absorptivity for each of these configurations. Interfering optical reflections at the interface of different layers cause a dependence on the wavelength and the exact layer structure. Rad-Pro utilizes universally accepted and extensively calibrated models to predict the directional, spectral, and temperature dependence of radiative properties for multilayer structures consisting of materials like silicon (doped/undoped), silicon dioxide, silicon nitride, and polysilicon.

Figures 8 and 9 show the plots of normalized emissivity and absorptivity, respectively, as a function of temperature, for the different layer structures. In the figures, 0 represents nominal anneal temperature for the 45nm technology used for this project, and other temperature points are expressed as difference from nominal. We observed that the variation of these optical properties with temperature can be modeled accurately by using linear functions of temperature. So, we modeled them in the form $a(bT+c)$, where a is the value of the property at nominal anneal temperature, and b and c are coefficients modeling the linear dependence on temperature. For example, the emissivity for STI is expressed as $\varepsilon_{STI}(b_{STI}T + c_{STI})$, where ε_{STI} is the emissivity for STI at nominal anneal temperature. Figures 7 and 8 also show the corresponding linear fit for each of the functions.

We use Calibre layout verification tool [14] to calculate relative densities of different layer types for each node point, in a rectangle with sides Δx, and Δy, centered at the node point. These density values are used to calculate density based weighted average of the linear fit coefficients, to yield final average temperature-dependent functions for absorptivity, and emissivity. Finally, we use MATLAB to solve for these simultaneous non-linear equations to yield chip level temperature maps for local anneal temperature. Local anneal temperature at any point can be determined by interpolating between the values at node points for the rectangle on the grid that contains the point. These temperature values are used in conjunction with TCAD based models for accurate performance and leakage analysis, by changing the delay/leakage based on the position of the gate in the layout. Next section discusses the result of such an analysis for some test chips.

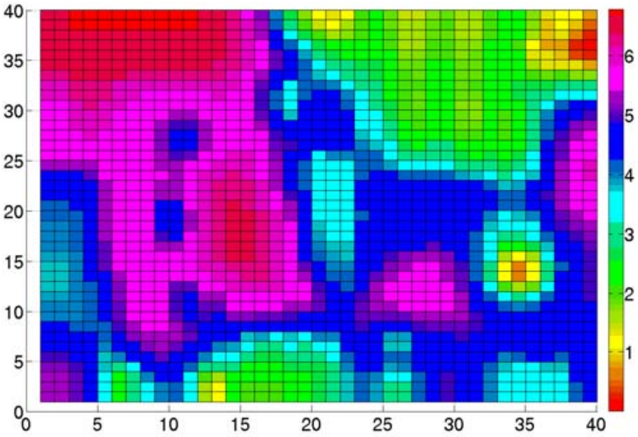

Figure 11. Ion map for the 45nm chip.

4. Experimental Results

In order to demonstrate the importance of anneal temperature variation aware analysis, we ran the flow described in the previous section for a 45nm experimental test chip. There are 40X40 rectangular grid cells of equal size located on the top surface of the chip, and a 40X40 temperature distribution map is calculated. Calibre was used for processing the layout, and obtaining the relative densities of different layer structures. MATLAB was used for solving the non-linear system of equations, and a trust-region based technique was utilized, which solves a linear system of equations to find the search direction [15, 16].

Figure 10 shows the temperature map for the chip. The X and Y axes represent the grid points on the 40X40 grid, while temperature is plotted as a function of position on the grid. The temperature map shows the presence of two high temperature regions on the chip, which result in the two peaks. The difference between maximum and minimum local anneal temperature on the chip was found to be ~10.5°C. We observed temperature map relates well with the STI density distribution, because STI has the lowest reflectivity amongst all the layer structure. Lower reflectivity translates into higher absorptivity, and a high density of STI results in higher temperature in the region due to increased absorption of incident power. However, there are other long range effects related to characteristic thermal length, which make the correlation less exact.

In order to examine the electrical effects of the local anneal temperature, the TCAD level models for temperature dependence of Ion and Ioff were employed. For both Ion and Ioff, the values reported are the average values for PMOS and NMOS. Figure 11 shows the Ion map for the chip. Ion values are reported as the percentage deviation from the slowest location on the die. Deviation of up to ~6.8% from the slowest location are observed.

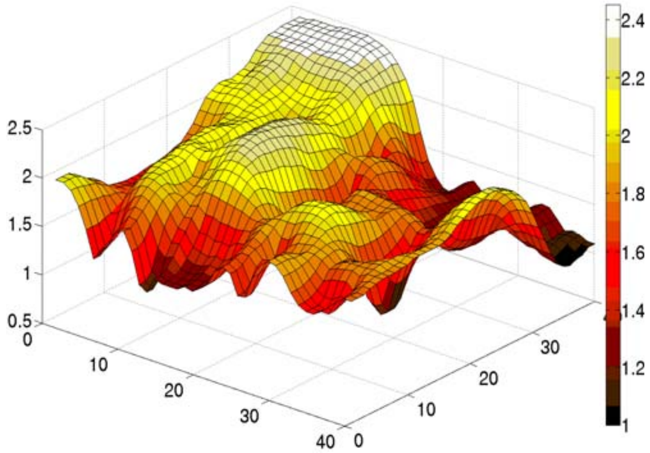

Figure 12. Ioff map for the 45nm chip.

The resulting deviation in inverter delay was found to be ~7.3%. A high deviation in performance/delay establishes the need for such a local anneal temperature aware performance analysis. Figure 12 shows the corresponding plot for Ioff. The effects observed here are substantial as well. The plot shows Ioff as a function of position on the grid, and all the values are normalized to the lowest leakage point in the die. We observe that device leakage at the fastest point (highest Ion) is ~2.45X higher than the slowest point.

To examine how the effect scales with technology, we performed the full chip thermal simulations for a 65nm test chip. We used a 40X40 rectangular grid for our analysis. Figure 13 shows the chip level temperature dustribution. The difference between maximum and minimum local anneal temperature on the chip was found to be ~8.5^{o}C, which is slightly smaller than the 45nm test case. The temperature distribution, again, shows a good correlation with STI density distribution. Although the magnitude of chip level temperature variation is smaller than the 45nm test case, it is high enough to cause a reasonable impact on performance and leakage.

5. Conclusion

In this work, we proposed a new local anneal temperature variation aware performance and leakage analysis framework which embodies transistor level models for anneal temperature sensitivity, to incorporate RTA induced temperature variation into traditional timing/leakage analysis. We solve for chip level anneal temperature distribution by dividing the wafer surface into rectangular grids, and employ TCAD based device level models for drive current (Ion) and leakage current (Ioff) dependence on anneal temperature variation, to capture the variation in device performance and leakage based on its position in the layout. Experimental results based on a 45nm test chip shows anneal temperature variations of up to 10.5^{o}C, which results in ~6.8% variation in device performance and 2.45X variation in device leakage across the chips. The corresponding variation in inverter delay was found to be ~7.3%, thereby establishing the importance of such a local anneal temperature variation aware performance/leakage analysis.

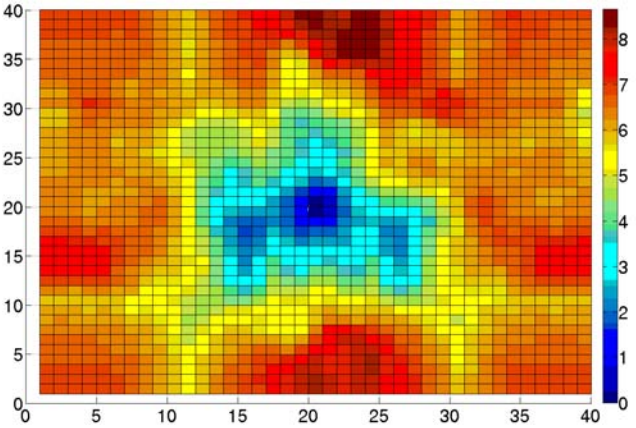

Figure 13. Local anneal temperature distribution for the 65nm chip.

References

[1] P. J. Timmans et al., “Rapid Thermal Processing,” in *Handbook of Semiconductoer Manufacturing Technology*, Y. Nishi and R. Doering (eds.), Marcel Dekker, Inc., New York (2000), pp. 201-286.

[2] R. B. MacKnight et al., “RTP application and technology options for the sub-45 nm nodes,” in *Proc. 12th IEEE International Conference on Advanced Thermal Processing of Semiconductors*, pp. 3-36, 2004.

[3] A. Agarwal, “Ultra-shallow junction formation using conventional ion implantation and rapid thermal annealing,” in *Proc. Conference on Ion Implantation Technology*, pp. 293-299, 2000.

[4] P. J. Timmans et al., “Challenges for ultra-shallow junction formation technologies beyond the 90 nm node,” in *Proc. 11th IEEE International Conference on Advanced Thermal Processing of Semiconductors*, pp. 17-33, 2003.

[5] Woo Sik Yoo et al., “Comparative study on implant anneal using single wafer furnace and lamp-based rapid thermal processor,” in *Proc. 9th IEEE International Conference on Advanced Thermal Processing of Semiconductors*, pp. 240-245, 2001.

[6] T. Feudel et al., “Junction scaling for next generation microprocessor technologies,” in *Proc. SEMI Technology Sessions, SEMICON Europa*, 2005.

[7] Y. H. Zhou et al., “A Monte Carlo Model for Predicting the Effective Emissivity of the Silicon Wafer in Rapid Thermal Processing Furnaces,” in *International J. Heat Mass Transfer, vol. 45*, pp. 1945-1949, 2002.

[8] S. K. Springer et al., “Modeling of Variation in Submicrometer CMOS ULSI Technologies,” in *IEEE Transactions on Electron Devices, vol. 53, no. 9*, pp. 2168-2178, 2006.

[9] I. Ahasan et al., “RTA-driven intra-die variations in stage delay, and parametric sensistivities for 65nm technology,” in *VLSI Symp. Tech. Digest*, pp. 170-171, 2006.

[10] M. Rabus et al., “Rapid thermal processing of silicon wafers with emissivity patterns,” in Journal of Electronic Materials, vol. 35, no. 5, pp. 877-891, 2006.

[11] Manual, Synopsys TSUPREM4, Version 2007.03.

[12] Manual, Synopsys MEDICI, Version 2007.03.

[13] B. J. Lee et al., “Rad-Pro: effective software for modeling radiative properties in rapid thermal processing,” in *Proc. 13th IEEE International Conference on Advanced Thermal Processing of Semiconductors*, pp. 275-281, 2005.

[14] Manual, Mentor Graphics CALIBRE, Version 2008.03.

[15] J. J. More et al., “Computing a trust region step,” SIAM Journal on Scientific and Statistical Computing, vol. 3, pp. 553-572, 1983.

[16] R. H. Byrd et al., “Approximate solution of the trust region problem by minimization over two-dimensional subspaces,” Mathematical Programming, vol. 40, pp. 247-263, 1988.