Compact Thermal Models for Estimation of Temperature-dependent Power/Performance in FinFET Technology

Aditya Bansal, Mesut Meterelliyoz, Siddharth Singh[*], Jung Hwan Choi, Jayathi Murthy[§], Kaushik Roy
School of Electrical and Computer Engineering, Purdue University,
[§]Department of Mechanical Engineering, Purdue University
West Lafayette, IN 47907
[*]Department of Electrical Engineering, Osmania University, India
Email: {bansal, mesut, choi56, jmurthy, kaushik}@purdue.edu

Abstract: With technology scaling, elevated temperatures caused by increased power density create a critical bottleneck modulating the circuit operation. With the advent of FinFET technologies, cooling of a circuit is becoming a bigger challenge because of the thick buried oxide inhibiting the heat flow to the heat sink and confined ultra-thin channel increasing the thermal resistivity. In this work, we propose compact thermal models to predict the temperature rise in FinFET structures. We develop cell-level compact thermal models for standard INV, NAND and NOR gates accounting for the heat transfer across the six faces of a cell. Temperature maps of benchmark circuits exhibit close correspondence with dynamic power maps because of confined regions of heat generation separated by low thermal conductivity material. It is illustrated that temperature-aware timing analysis is imperative, because of high inter-cell temperature gradient. Accurate prediction of temperature in the early phase of design cycle will give valuable estimation of power/performance/reliability of a circuit block and will guide in the design of more robust circuits.

I. Introduction

The need for higher performance in smaller area has always been the driving force for the semiconductor industry. Aggressive technology scaling has led to increase in transistor density on a chip and innovative device structures (UTB-SOI, FinFET, Tri-gate etc.) to achieve the desired performance. Increasing packing density has led to power density to become a critical bottleneck in the design of microelectronics. The local temperature rise can result in circuit malfunction and can also impact performance, power and reliability. For every 10°C increase in temperature, a MOSFET's drive current decreases approximately 4% and interconnect (Elmore) delay increases approximately 5% [1]. To avoid the increase in temperature (or to dissipate heat), a heat sink is integrated on the chip package. However, since the heat sink is away from the device layer, it is not very efficient in taking the heat directly away from the transistors.

Several architecture and circuit level techniques have been proposed to distribute the temperature uniformly over the chip and reduce the hot spots. At architecture level, several dynamic thermal management (DTM) schemes have been proposed including dynamic voltage scaling (DVS) [2], discrete frequency scaling (DFS), migrating computation (MC) [3] and dynamic clock throttling (DCT) [4] etc. These techniques depend on the accurate thermal modeling of circuit blocks. Conventionally, circuit blocks are modeled as heat sources with uniform temperature distribution inside the block. There are several thermal models available in the literature at different stages of design cycle of VLSI circuits. For example, a dynamic compact thermal model at microarchitecture level is presented in [3]. Grid-like thermal models for temperature-aware-design at arbitrary granularities are presented in [5]. Weiping et al. [6] modeled the dependence of power and performance on the temperature at microarchitecture level. However, thermal modeling at finer granularity level i.e., transistor level or logic gate level is required for more accurate estimation of local hot spots.

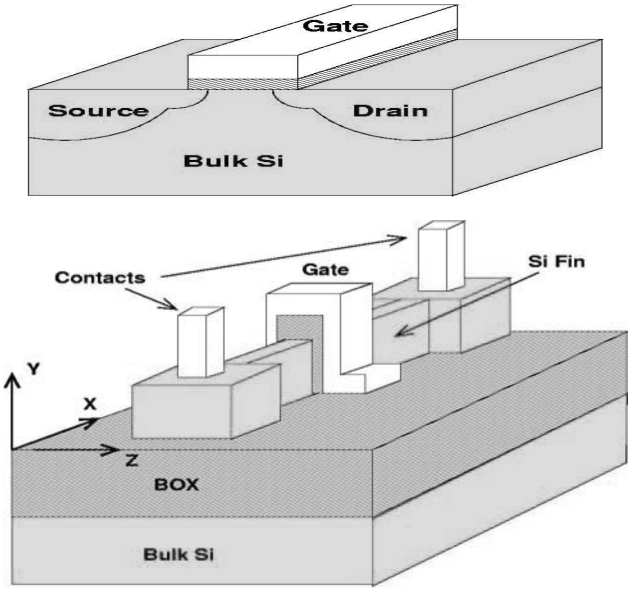

Fig. 1: Schematics of bulk-Si MOSFET and FinFET. In FinFET, gate surrounds the channel which is separated from the bulk-Si by thick BOX.

The increasing thermal problems are more aggravated in next generation transistor technologies such as Ultra-thin-body (UTB) Silicon-On-Insulator (SOI) MOSFETs and FinFETs. FinFETs have gate insulator all around the channel where heat is generated. This oxide layer results in less efficient dissipation of heat compared to bulk-MOSFETs where heat mainly dissipates through the substrate [7]. Eric et al. [8] [9] discuss the impact of confined dimensions and complicated geometries on self-heating in these devices. Ultra-thin silicon body improves the device scalability by reducing short-channel-effect, higher on-current and near ideal sub-threshold slope. However, the thermal resistivity of the thin active region is higher compared to thick active region in bulk-MOSFETs. Also, because of increased surface

0-7803-9451-8/06/$20.00 ©2006 IEEE.

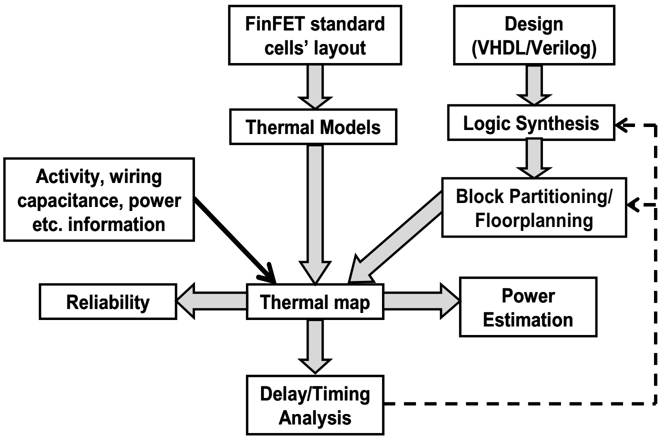
FinFET standard cells' layout
Design (VHDL/Verilog)
Thermal Models
Logic Synthesis
Activity, wiring capacitance, power etc. information
Block Partitioning/ Floorplanning
Reliability
Thermal map
Power Estimation
Delay/Timing Analysis

Fig. 2: Design flow for temperature based power/ performance/reliability estimation.

to volume ratio in UTB-SOI devices, thermal boundary resistances of various materials' further impede the heat flow [9]. Fig.1 shows the schematics of bulk- MOSFET and FinFET devices. In bulk-MOSFETs, heat generated in device layer as well as interconnect layers is mainly dissipated through the heat sink integrated with bulk substrate silicon. However, in FinFETs, device layer is separated from bulk-Si substrate by thick buried oxide layer (which has two orders of magnitude lower thermal conductivity than silicon [10]). Therefore, heat dissipation is mainly through the contact holes and interconnects. This elevates the local temperature rise in FinFET circuits. Because of temperature dependence of critical electrical parameters, it is becoming increasingly clear that electro-thermal co-design of transistors and circuits is necessary to arrive at optimal power and performance. However, a typical chip may have hundreds of millions of transistors, and a direct thermal simulation of such a collection of devices is all but impossible with current computing power. Instead, in this paper, we develop compact thermal models for circuit components used in cell-level electrical models in circuit design. Detailed computational models for three cell-level components – NAND, NOR and INV – are created based on Fourier theory.

Fig. 2 shows the design flow adopted in this work for the estimation of temperature-dependent power/performance/ reliability in FinFET based circuits. Compact thermal models are generated for standard cell layouts. The temperature information for each standard cell is associated with its heat generation and heat flow across the boundaries to the neighboring cells. Note that temperature of a cell is dependent on the heat generation in the transistors and temperatures of neighboring cells. Hence, real temperature can not be determined till a floorplan is generated. These models are used in generating the temperature profile of a circuit for a given floorplan. Wiring capacitance is accounted for as lumped capacitive load at the output of driving gate. The temperature map is used to estimate the increase in power dissipation, deterioration in performance and lifetime reliability because of self-heating and local temperature rise. This estimation can be primarily used in two ways for robust circuit design: (1) It can be used to define an appropriate cost function to develop thermally-aware placement and optimization strategies for VLSI circuits, (2) Temperature map can be used for better delay analysis of the critical paths in early phase of the design cycle.

In this work, we are targeting the problem of heat dissipation in predictive 28nm ITRS [11] specified technology node for FinFET circuits. In section 2, we discuss the impact of local temperature rise on the power and performance of the standard cells. In section 3, we develop the compact thermal models for standard logic cells designed using NMOS and PMOS FinFETs. The joule heating in devices is simulated for the worst case input pattern (each transistor switching at every clock cycle) using Taurus Device Simulator [12]. Switching activity dependent temperature rise in a standard cell is modeled based on three-dimensional thermal resistances of the cell. In section 4, we generate the temperature maps of benchmark circuits and discuss the local temperature rise in FinFET circuits followed by conclusions in section 5.

II. Performance and Power Estimation

The temperature rise in a transistor affects its electrical characteristics. Mobility and sub-threshold slope degrade with the increase in temperature resulting in reduced on-current and hence, increased delay. Also, static leakage increases with temperature because of increase in sub-threshold leakage. In scaled technology generations, the static power is becoming a significant component of the total power dissipation, especially in the low activity sections of a chip like memory. Moreover, reliability is strongly dependent on the temperature. Increasing the temperature exponentially decreases the lifetime of a circuit. A first order model of mean-time-to-failure (MTF) can be given by Arrhenius equation:

$$MTF = MTF_0 \exp(E_a/k_BT) \qquad (1)$$

exhibiting the exponential deterioration in reliability with temperature. We quantitatively analyze the temperature dependence of delay and power in a 2-input NOR gate. Similar analysis has been done for INV and 2-input NAND gates, however, results are omitted for brevity.

A. Delay dependence on temperature

With the increase in temperature, mobility degrades in transistors. This effect is more dominant in bulk MOSFETs because of heavily doped body used to reduce short-channel-effect. However, in fully-depleted double gate SOI devices, body is left undoped resulting in less deterioration of mobility with temperature. Mobility degradation reduces on-current, however, with increase in temperature, threshold voltage decreases resulting in improved on-current. These two effects counter each other and the net variation in on-current is dependent on the relative sensitivities of mobility and threshold voltage to the temperature. To analyze the impact of temperature on the intrinsic delay, we simulated a three NOR stage ring oscillator with inputs of each NOR gate tied together. Fig. 3 shows the intrinsic delay increase with temperature. It can be seen that intrinsic delay increases

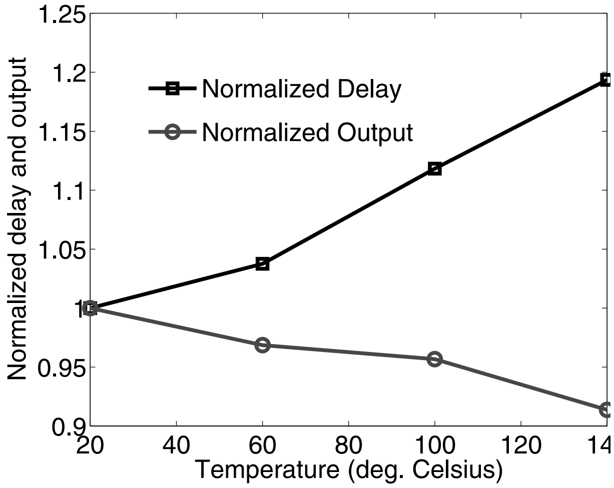

Fig. 3: Normalized delay and output swing of 3 NOR stage ring oscillator with temperature.

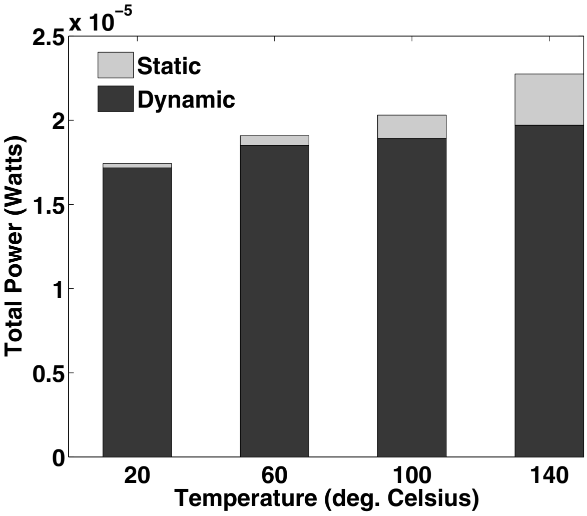

Fig. 4: Total power dissipation in NOR cell with temperature.

approximately 20% for 120°C rise in cell temperature. This increase in intrinsic delay coupled with increase in Elmore delay of interconnects because of increased electrical resistivity can pose serious challenges to circuit designers demanding more rigorous timing analysis. One more observation is the decrease in output swing (Fig. 3) because of increase in sub-threshold slope and reduced I_{on}/I_{off} ratio. Output swing decreases 9% for 120°C rise in temperature. This will result in reduced noise margins in cascaded logic gates at high temperatures.

B. Power dependence on temperature

Power dissipation in a circuit depends on its operating mode and can be given by

$$P_{total} = P_{static} + P_{dynamic} \tag{2}$$

Static power dissipation (P_{static}) is due to the leakage currents – sub-threshold leakage, gate direct tunneling leakage and reverse-biased junction band-to-band tunneling leakage – in off-state of a transistor [13]. Sub-threshold leakage increases exponentially with the decrease in threshold voltage and is given by,

$$I_{sub} = I_o e^{q(V_{gs}-V_t)/mk_BT}(1 - e^{-qV_{ds}/k_BT}) \tag{3}$$

This results in 12X increase in static leakage for the temperature rise of 120°C in a NOR gate (Fig. 4).

Dynamic power dissipation ($P_{dynamic}$) is mainly due to:

(1) charging/discharging of the capacitances at the output and internal nodes of a cell ($P_{cap,dyn}$).

(2) short-circuit power due to direct current path between the power supply and ground ($P_{sc,dyn}$).

The switching of the capacitances at the output and internal nodes of a cell depends on the input pattern. Average dynamic power dissipation can be obtained by considering the average number of transitions per clock cycle. The dynamic power can be given by,

$$P_{cap,dyn} = 0.5 \times V_{DD}^2 \times f \times \left[\left(C_{load} \times \alpha_{out} \right) + \sum (C_i \times \alpha_i) \right] \tag{4}$$

where, V_{DD} is the supply voltage and f is the clock frequency. C_{load} includes the output load capacitance and the device capacitances of transistors connected at the output node and α_{out} is the average number of output transitions in a clock cycle. C_i is the internal node capacitance of the i^{th} node and α_i is the average number of switching transitions per clock cycle at the i^{th} node.

Short-circuit power is due to direct path of current flow between the power supply and ground during output transition. Because of the non-zero rise/fall times of an input signal, pull-down and pull-up networks can be simultaneously conducting in a CMOS circuit resulting in a direct current path between power supply and ground. With the increase in temperature, sub-threshold slope of the transistors increases and I_{on}/I_{off} ratio decreases. This results in increased short circuit power. Fig. 4 shows the dynamic and static power dissipation in a NOR gate. Dynamic power dissipation is calculated by integrating the currents in a 3-stage ring oscillator. It can be seen that dynamic power dissipation slightly increases with temperature mainly because of increase in short-circuit power.

III. Compact Thermal Models

A. Detailed Component Model

Detailed computational models for three cell-level components – NAND, NOR and INV are created based on Fourier theory. A typical NOR gate structure is shown in Fig. 5, including two PMOS and two NMOS FinFETs at the 28 nm node. The gates and fins are also shown. The simulation includes portions of the metallic contacts, but not all the metallization layers. These structures are enclosed in a domain of size 504 nm x 300 nm x 539 nm and are located on the y=150 nm plane. The rest of the interior volume is filled with SiO_2. In this orientation, the metallization layers would lie above the y=300 nm boundary, while the heat sink would lie below the y=0 boundary. Other details for the three components are shown in Table 1. The gate material is assumed to be aluminum, as are the metallic contacts, and

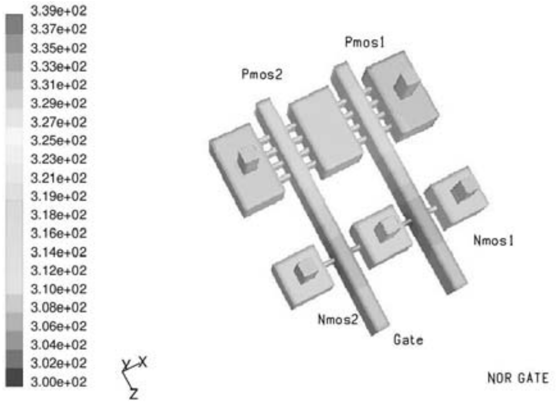

Fig. 5: Temperature field in NOR gate on the y=150 nm plane. Two PMOS and two NMOS FinFETs at the 28 nm technology node are shown. Heat generation in the fingers causes a temperature rise of 39°C above ambient.

standard temperature-independent thermal properties for bulk Si, SiO_2 and aluminum are assumed.

The volumetric heat generation due to self-heating is computed using the TAURUS device simulator [12]. The electrical simulation is 2D, in the x-z plane in Fig. 5, and yields the heat source **J.E** (x, z) for a typical fin in the PMOS or NMOS device. Thermal transport in the geometry is necessarily 3D with the primary direction for heat transfer being perpendicular to the x-z plane. The spatial (x, z) distribution of the heat generation is assumed to hold throughout the y-depth of the fins. The FLUENT CFD software [14] is used to compute the temperature field using the unstructured mesh technique described in [15].

B. Compact Model Generation

The generation of the compact model exploits the fact that the Fourier conduction equation with constant thermal properties is a linear elliptic boundary value problem. Thus, the temperature at any location (x_1, y_1, z_1), or indeed, the average temperature of the domain, can be uniquely written as:

$$T\left(x_1,y_1,z_1\right)=\sum_{f=1}^{6}a_f\left(x_1,y_1,z_1\right)T_f+a_0\left(x_1,y_1,z_1\right)\alpha q \quad (5)$$

where the coefficients a_f are unique to the spatial location (x_1 ,y_1,z_1), and determine the influence of the six boundary temperatures of the cuboidal domain on T(x_1, y_1, z_1). The coefficient a_0 quantifies the influence of the heat generation rate q for a given activity α. The coefficients a_f have the property:

$$\sum_{f=1}^{6}a_f=1 \quad (6)$$

By the same token, the heat transfer rates q_{bj} out of each of the six boundaries of the domain may also be written as:

$$q_{bj}=\sum_{f=1}^{6}b_{fj}T_f+b_{0j}\alpha q \quad \text{j=1,2,...6} \quad (7)$$

For a given geometry, materials properties, and spatial pattern of heat generation, the coefficients a_f, a_0, b_{fj} and b_{0j} are uniquely determined, and constitute the compact model of the NOR, NAND or INV component. Seven temperature calculations are done using seven different sets of boundary temperatures to obtain a total of seven coefficient values (six a_f's and a_0). The same seven runs are also sufficient to determine seven values (b_{fj} and b_{0j}) for each of the six boundary face heat transfer rates q_{bj}, j=1,2,…6 . A typical compact model for a NOR gate is present in Table 2. Here, the average component temperature is related to the six boundary temperatures of the computational domain and the heat generation rate; similarly, the heat transfer rates at the six boundary faces. This compact model will recover *exactly the same* average temperature and boundary heat transfer as the detailed model for any set of Dirichlet boundary conditions on the boundaries.

Table I: Parameters for Thermal Simulations

Comp-onent	Devices	Domain Size	Mesh Size (Hexah-edra)	Number of Fins	Heat Generation Per Fin (W)	Net Heat Generation in Device (W)
NOR	2 PMOS, 2 NMOS	504 nm x 300 nm x 539 nm	267,376	PMOS1 : 4 PMOS2: 4 NMOS1: 1 NMOS2: 1	PMOS1 : 4.4399×10^{-6} PMOS2: 5.211×10^{-6} NMOS1: 8.2332×10^{-5} NMOS2: 8.2332×10^{-5}	2.0326×10^{-4}
NAND	2 NMOS 2 PMOS	504 nm x 300 nm x 406 nm	146,260	PMOS1 : 2 PMOS2: 2 NMOS1: 2 NMOS2: 2	PMOS1 : 3.1178×10^{-5} PMOS2: 2.1156×10^{-5} NMOS1: 8.4242×10^{-6} NMOS2: 8.4242×10^{-6}	1.3836×10^{-4}
INV	1 PMOS 1 NMOS	336 nm x 300 nm x 399 nm	228,214	PMOS1 : 2 NMOS1: 1	PMOS1 : 9.496×10^{-7} NMOS1: 3.4595×10^{-6}	5.3589×10^{-6}

Table II: Compact thermal model for NOR gate

$T_{avg} = 0.1258T_{left} + 0.1256T_{right} + 0.2326T_{bottom} + 0.2747T_{top} + 0.1228T_{back} + 0.1185T_{front} + 3.5039x10^{4}\alpha q$
$q_{left} = 7.2206x10^{-6}T_{left} - 6.8573x10^{-8}\ T_{right} - 2.2603x10^{-6}\ T_{bottom} - 2.5307x10^{-6}\ T_{top} - 1.1725x10^{-6}\ T_{back} - 1.1885x10^{-6}\ T_{front} - 0.1141\ \alpha q$
$q_{right} = -6.8573x10^{-8}T_{left} + 7.2188x10^{-6}\ T_{right} - 2.2603x10^{-6}\ T_{bottom} - 2.5301x10^{-6}\ T_{top} - 1.1719x10^{-6}\ T_{back} - 1.1879x10^{-6}\ T_{front} - 0.1119\ \alpha q$
$q_{bottom} = -2.2603x10^{-6}T_{left} - 2.2603x10^{-6}\ T_{right} + 9.2783x10^{-6}\ T_{bottom} - 5.0752x10^{-7}\ T_{top} - 2.0967x10^{-6}\ T_{back} - 2.1534x10^{-6}\ T_{front} - 0.1168\ \alpha q$
$q_{top} = -2.5307x10^{-6}T_{left} - 2.5301x10^{-6}\ T_{right} - 5.0752x10^{-7}\ T_{bottom} + 1.0289x10^{-5}\ T_{top} - 2.3629x10^{-6}\ T_{back} - 2.3574x10^{-6}\ T_{front} - 0.3877\ \alpha q$
$q_{back} = -1.1725x10^{-6}T_{left} - 1.1719x10^{-6}\ T_{right} - 2.0967x10^{-6}\ T_{bottom} - 2.3629x10^{-6}\ T_{top} + 6.9145x10^{-6}\ T_{back} - 1.1047x10^{-7}\ T_{front} - 0.1264\ \alpha q$
$q_{front} = -1.1885x10^{-6}T_{left} - 1.1879x10^{-6}\ T_{right} - 2.1534x10^{-6}\ T_{bottom} - 2.3574x10^{-6}\ T_{top} - 1.1047x10^{-7}\ T_{back} + 6.9977x10^{-6}\ T_{front} - 0.1430\ \alpha q$
Left ($x=0$); right ($x=x_{max}$); bottom($y=0$); top ($y=y_{max}$); back ($z=0$); front ($z=z_{max}$)

IV. Results and Discussions

Once the average temperature and boundary heat transfer dependence of a logic cell on the boundary temperatures of the cuboid are known, the next step is cell placement and the generation of the thermal map of a circuit block. Though the primary heat flow direction is perpendicular to the plane in which the logic cells are placed, there is sufficient in-plane thermal non-uniformity that the thermal transport in all three coordinate directions must be considered. The logic cells are arranged in planar mesh of cuboidal cells, with those cells not containing logic elements assumed to contain SiO_2. Convective heat transfer boundary conditions are posed on all external boundaries, with the heat transfer coefficients being chosen to correctly model the thermal resistance due to metallization layers, the wafer, as well as lateral losses to other circuit blocks. By enforcing continuity of heat transfer rate at logic cell cuboid faces, equations for the face temperatures are found. These, in turn, are used to evaluate the average logic cell temperature. Heat generation and hence local temperature rise in a cell depends on the input data pattern. Therefore, to obtain the worst cast temperature map, an exhaustive set of test patterns need to be applied. However, to reduce the computational complexity, we obtain the average switching activity of each cell in a circuit block for a large set of random input patterns.

Fig. 6 shows the dynamic power and temperature distribution in the layouts of *alu4* and *x3* (MCNC'91 [16]) benchmark circuits. Circuits have been modified to use standard cells. Floorplans are generated to arrange cells in a planar grid of 30x30. Each cell volume is either occupied by a standard cell – NAND, NOR, INV – or filled with SiO_2 insulator. The maximum temperature difference inside the circuit blocks is 20°C. It can be seen that in FinFET based circuits, the temperature distribution closely corresponds to the dynamic power map (Fig.6), unlike bulk MOSFETs where hot spot region distributes over several cells. This is attributed to the confined silicon channels and lack of common high conductivity bulk silicon under the device layer.

In FinFET circuits, lateral heat flow is mainly through interconnects because of the lack of common low thermal resistivity substrate. This can result in large temperature differences between the neighboring cells. Power and ground lines shared by the adjacent cells help in heat flow and temperature distribution to some extent. Because of high inter-cell temperature gradients, it's imperative to perform temperature-dependent timing analysis at the granularity of gate level in FinFET circuits. With this estimation, some circuit level techniques, such as sizing, can be employed to prevent the failures.

V. Conclusions

In this work, we propose gate level compact thermal models for estimating temperature rise in FinFET circuits. The effect of low thermal conductivity buried oxide, confined ultra-thin channel and thermal boundary resistances of different materials' on temperature rise is taken into account. The floorplans of benchmark circuits show close correspondence between the temperature maps and dynamic power maps. This can be attributed to the confined channels where heat is generated and lack of high thermal conductivity material (bulk silicon) under the device layer impeding the lateral heat distribution. Results show that 120°C rise in temperature can result in 20% increase in intrinsic delay of a 2-input NOR gate and 12X increase in leakage power dissipation. The proposed thermal models can be used in estimating temperature rise in early phase of design cycle for proper timing analysis. The models can also be integrated in floorplanning algorithms to remove hot spots.

References

[1] C.-H. Tsai et al., "Standard Cell Placement for Even On-Chip Thermal Distribution," *Proc. of Int. Sym. on Phys. Design,* 1999, pp. 179-184.

[2] D. Brooks et al., "Dynamic Thermal Management for High-Performance Microprocessors," *Proc. 7th Int. Sym. High-Perf. Comp. Arch.*, 2001, pp. 171-182.

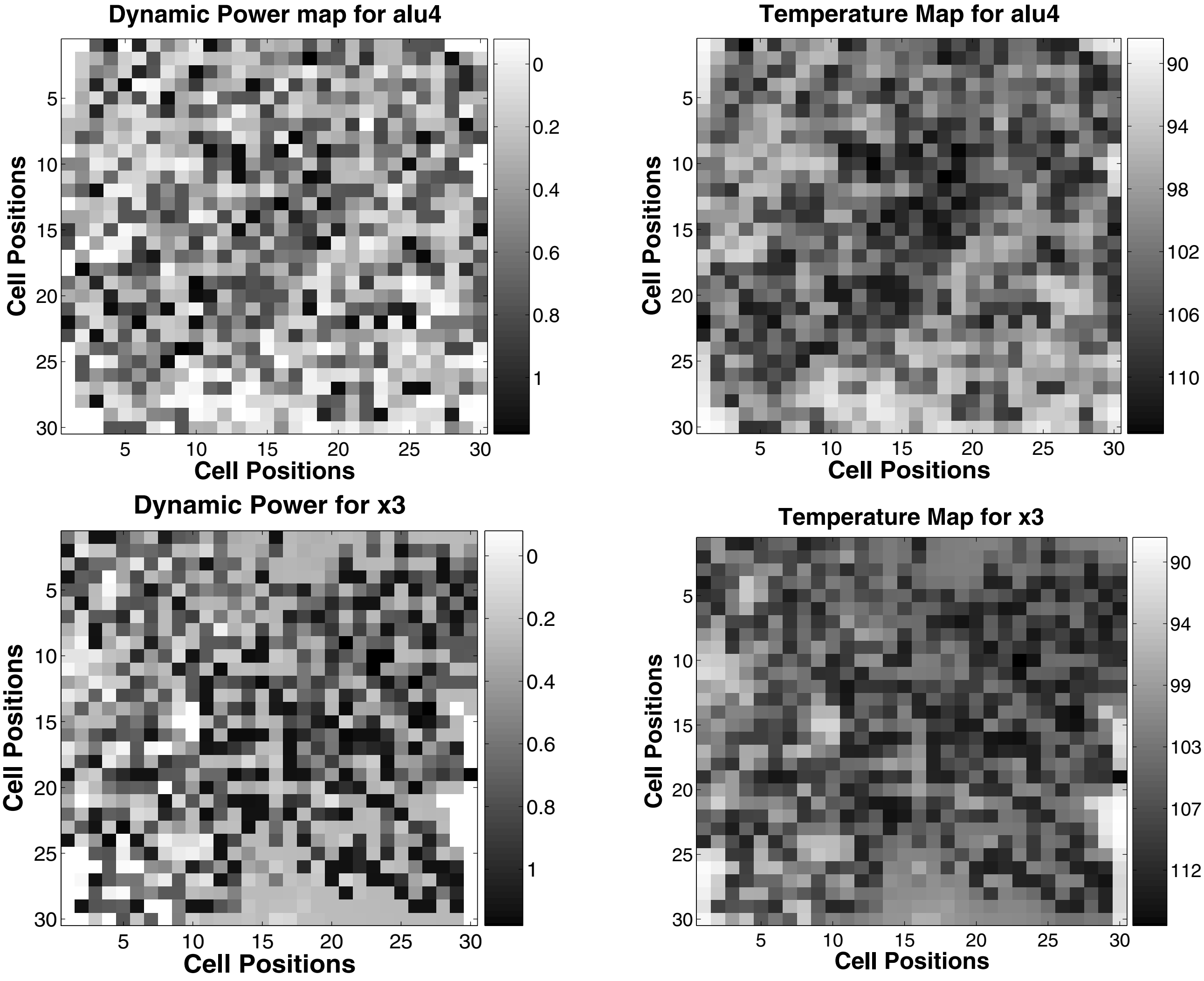

Fig. 6 Dynamic power and temperature maps for *alu4* and *x3* benchmark circuits. x-y axes represent the positions of standard cells. Dynamic power is normalized to maximum, where zero value corresponds to empty cells filled with oxide.

[3] K. Skadron et al., "Temperature-Aware Microarchitecture: Modeling and Implementation" *ACM Transactions on Architecture and Code Optimization*, Vol. 1, No. 1, March 2004, pp. 94-125.

[4] W. R. Daasch et al., "Design of VLSI CMOS Circuits Under Thermal Constraint," *IEEE Trans. on Ckts. and Sys. – II*, Aug. 2002, pp. 589-593.

[5] W. Huang et al., "Compact Thermal Modeling for Temperature-Aware Design," *DAC*, 2004, pp. 878-883.

[6] W. Liao et al., "Temperature and Supply Voltage Aware Performance and Power Modeling at Microarchitecture Level," *IEEE Trans. on Computer-Aided Design of Integrated Circuit and Systems,* Vol. 24, No. 7, July 2005.

[7] M. Berger et al., "Estimation of Heat Transfer in SOI MOSFET's," *IEEE Trans. on Elec. Dev.*, 1991, pp. 871-875.

[8] E. Pop et al., "Thermal Analysis of Ultra-thin Body Device Scaling," *IEDM*, 2003, pp. 883-886.

[9] E. Pop et al., "Thermal Phenomena in Nanoscale Transistors," *ITherm*, 2004, pp. 1-7

[10] L. T. Su et al., "Measurement and Modeling of Self-Heating in SOI NMOSFETs," *IEEE Trans. on Elec. Dev,* 1994, pp. 69-75.

[11] The International Technology Roadmap for Semiconductors, *Semiconductor Industry Assoc*, 2004update.

[12] Taurus Device Simulator v2004.09, *Synopsys Inc*.

[13] K. Roy et al., "Leakage Current Mechanisms and Leakage Reduction Techniques in Deep-Submicrometer CMOS Circuits," *Proc. of IEEE*, 2003, pp. 305-327.

[14] Fluent User Manual, Fluent Inc., Lebanon, NH03766, 2005.

[15] S. R. Mathur and J. Y. Murthy, *Advances in Numerical Heat Transfer*. Taylor and Francis, 2000, vol. 2, pp. 37-67.

[16] S. Yang, *Logic Synthesis and Optimization Benchmarks User Guide Version 3.0*. MCNC, Research Triangle Park, NC, Jan. 1991.

.