Leakage Power Dependent Temperature Estimation to Predict Thermal Runaway in FinFET Circuits

Jung Hwan Choi, Aditya Bansal, Mesut Meterelliyoz, Jayathi Murthy[++], Kaushik Roy

Electrical and Computer Engineering; [++]Mechanical Engineering
Purdue University
West Lafayette, IN 47907, USA

{choi56, bansal, mesut, jmurthy, kaushik}@ecn.purdue.edu

ABSTRACT

In this work we propose a methodology to self-consistently solve leakage power with temperature to predict thermal runaway. We target 28nm FinFET based circuits as they are more prone to thermal runaway compared to bulk-MOSFETs. We generate thermal models for logic cells to self-consistently determine the temperature map of a circuit block. Our proposed condition for thermal runaway shows the design trade off between the primary input (PI) activity of a circuit block, sub-threshold leakage at the room temperature and the thermal resistance of the package. We show that in FinFET circuits, thermal runaway can occur at the ITRS specified sub-threshold leakage (150nA/μm, high-performance) for a nominal PI activity of 0.5 and typical package thermal resistance.

1. INTRODUCTION

Conventional bulk-MOSFET structure suffers from scalability in sub-100nm nodes because of deteriorating device electrostatics resulting in increased short channel effect (SCE). To improve electrical characteristics, several innovative device structures such as ultra-thin-body SOI and FinFET [1] have been proposed. FinFETs have emerged as the best candidate because of better scalability and ease of fabrication. However, FinFETs have confined channel, surrounded by silicon dioxide, which has lower thermal conductivity compared to bulk silicon [2]. This results in increased self-heating and aggravated thermal issues [3]. Furthermore, the ITRS [4] predicted sub-threshold leakage ceiling for FinFET (double-gate FET for high performance) is comparable to planar bulk MOSFETs. Therefore, these devices almost equally suffer from increasing leakage power in total power consumption. Under high frequency operation, temperatures rise due to large active power consumption. The high temperature increases the sub-threshold leakage (which is strong function of temperature), further increasing temperature. If heat cannot be dissipated effectively, a positive feedback between leakage power and temperature can result in thermal runaway. To predict thermal runaway, it is important to account for all the components of power dissipation self-consistently with respect to temperature.

In earlier work, researchers have predicted thermal runaway by calculating the junction temperature (T_j) of the silicon substrate based on the ambient temperature (T_a) and the thermal resistances of the package [5]. Temperature of the silicon substrate is calculated and successively updated based on the total power dissipation in a chip and thermal resistance from junction to air (θ_{ja}). T_j can be given by [6]

$$T_j = T_a + P \cdot \theta_{ja} \tag{1}$$

However, this approach cannot be used for FinFET circuits because device channel regions are separated by silicon dioxide and hence, temperatures at finer granularity level need to be considered.

In this paper we propose a methodology to solve total power self-consistently with temperature to predict thermal runaway in FinFET based circuits. In particular in this work

- We employ the gate level (INV, NAND and NOR) thermal models to generate thermal map of a circuit block. Cell level modeling accounts for the lateral heat flow between the neighboring cells along with vertical heat flow to the heat sink.
- We self-consistently solve leakage power (sub-threshold) with temperature to calculate static temperature rise along with activity dependent dynamic power.
- Our analysis shows that for a nominal PI activity of 0.5 and package thermal resistance of 0.7K/W, thermal runaway can occur at the ITRS [4] specified sub-threshold leakage of 150nA/μm at 28nm technology node in FinFET circuits.
- Since limiting sub-threshold leakage limits on-current, we show that for high performance circuits, package quality should be improved i.e., package thermal resistance should be reduced to increase the maximum affordable sub-threshold leakage.

2. TEMPERATURE DEPENDENT POWER ESTIMATION

Power dissipation of a logic gate consists of dynamic power and static power. While dynamic power consumption is weakly coupled with temperature variation, static power consumption is a strong function of temperature. For estimation of temperature distribution consistent with static power consumption, temperature-dependent static power model is necessary.

2.1 Static Power

The major components of static power are sub-threshold leakage, gate direct tunneling leakage and junction band-to-band tunneling leakage [7]. In FinFET devices, junction leakage is not as significant as in bulk MOSFET since halo doping is not used for

ICCAD'06, November 5-9, 2006, San Jose, CA
Copyright 2006 ACM 1-59593-389-1/06/0011...$5.00

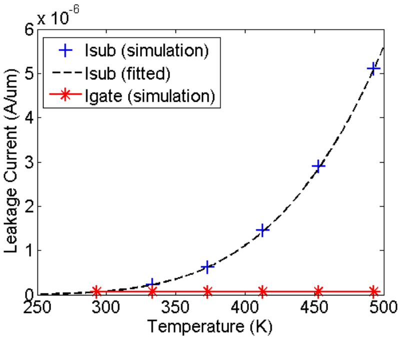


Figure 1. Leakage current components varying with temperature for nMOS FinFET transistor.

FinFETs. Therefore, we do not consider this leakage component in this work. Static power consumption can be given as

$$P_{static} = V_{DD} \cdot I_{static} = V_{DD} \cdot (I_{gate} + I_{sub}) \qquad (2)$$

where I_{gate} is gate direct tunneling leakage and I_{sub} represents sub-threshold leakage. While dynamic power is determined by switching activity of output, static current is determined by input signal states [8]. The average static power consumption of 2-input NAND, for example, is calculated as

$$P_{static} = V_{DD} \sum_{i,j=0,1} p_{ij} (I_{gate,ij} + I_{sub,ij}) \qquad (3)$$

where p_{ij} is the probability that the NAND gate has ij as an input vector and $I_{gate,ij}$ and $I_{sub,ij}$ are, respectively, gate tunneling leakage current and sub-threshold leakage current corresponding to input vector ij. For different input conditions, we simulated gate leakage current and sub-threshold leakage current using TAURUS device simulator [9].

Fig. 1 shows temperature dependencies of different components of leakage current. Although both sub-threshold leakage current (I_{sub}) and gate leakage current (I_{gate}) are known to vary with temperature, I_{sub} is by far more sensitive to temperature variation. At room temperature, I_{gate} is comparable to I_{sub}, however, at high temperatures sub-threshold leakage power becomes the dominant component of total static power consumption and the contribution of gate leakage becomes negligible. In this work, we assume that gate leakage current does not change with temperature.

Sub-threshold leakage current is expressed as [10]

$$I_{sub} \approx \mu_{eff} C_{ox} \frac{W}{L} (m-1) (\frac{kT}{q})^2 e^{-qV_t / mkT} \qquad (4)$$

where m is the body-effect coefficient. The temperature dependency of I_{sub} is dominated by $exp(-qV_t/mkT)$ term since T^2 is compensated by $\mu_{eff} \propto T^{-3/2}$. For sub-threshold leakage power estimation, we can expect that I_{sub} is suitably approximated with a fitting function $A \cdot exp(-B/T)$ where A and B are constants. Constants A and B are obtained from sub-threshold leakage values obtained by TAURUS simulation under several temperature conditions. As shown in Fig. 1, the fitting function (the dashed line in the figure) is well matched with simulated values for different temperatures.

2.2 Dynamic Power

Average dynamic power consumption is given by

$$P_{dynamic} = \frac{1}{2} C_L V_{DD}^2 \cdot \alpha \cdot f \qquad (5)$$

where C_L is switching capacitance, α is switching activity of output node, and f is the operating frequency of system. The switching capacitance C_L consists of three components:

$$C_L = C_{diff} + C_{wire} + \sum_i C_{gate,i} \qquad (6)$$

where C_{diff} is internal diffusion capacitances in this gate, $C_{gate,i}$ represents input gate capacitance of a fan-out gate, and C_{wire} is wire capacitance of output node. C_{diff} and C_{gate} are computed using TAURUS simulator and C_{wire} is calculated using wire length and wire capacitance per unit length. Wire length is estimated during cell placement step and wire capacitance per unit length is obtained from PTM [11]. The switching activity α of each gate is evaluated from the given input patterns. Since temperature has little impact on the switching activity and the load capacitance, dynamic power does not change much with temperature. We assume that dynamic power is insensitive to temperature in this work.

3. THERMAL MODELS FOR TEMPERATURE ESTIMATION

To generate the thermal profile of a circuit block, we create detailed computational models for three cell-level components - NAND, NOR and INV. Our models include buried oxide, fins, gates and portions of the metallic contacts, but not all the metallization layers. FinFET devices are placed on top of buried oxide of 150nm thickness and the gate material and the metallic contacts are assumed to be aluminum. Standard temperature-independent thermal properties are assumed for bulk Si, SiO_2 and aluminum.

3.1 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, y, z)*, or indeed, the average temperature of the domain, can be uniquely written as:

$$T(x,y,z) = \sum_{f=1}^{6} a_f(x,y,z) T_f + a_0(x,y,z) \cdot q \qquad (7)$$

where the coefficients a_f determine the influence of the six boundary temperatures (T_f) of the cuboidal domain on $T(x, y, z)$ and a_0 quantifies the influence of the heat generation rate q in cell.

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:

$$q_{bj} = \sum_{f=1}^{6} b_{fj} T_f + b_{0j} q \quad (j = 1,2,...,6) \qquad (8)$$

where b_{fj}'s are the coefficients relating heat transfer rate at face j to temperature of face f and b_{0j} is the relative heat transfer rate at face j to heat generation rate q in the cell. 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 components.

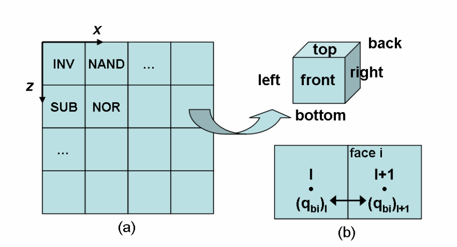


Figure 2. (a) Arrangement of cells in floor plan showing inverter (INV), NAND and NOR gates as well as substrate (SUB) cells. (b) Heat balance at face i between cell I and I+1.

3.2 Thermal Map Generation

Once compact thermal models have been created for the three gates under consideration, the next step is to include these in a cell-based representation of the floor plan. The logic cells are arranged in a planar mesh of cuboidal cells in the x-z plane, as shown in Fig. 2(a), with those cells not containing logic elements assumed to contain SiO_2 and labeled substrate (SUB). The discrete equation for the cell temperature is given by Eq. (7) which relates the cell average temperature to the temperatures on the six faces. And the face temperatures are found by enforcing the continuity of heat transfer rate at each cell face. For a face *i* shared by cells I and I+1 shown in Fig. 2(b),

$$(q_{bi})_I + (q_{bi})_{I+1} = 0 \tag{9}$$

Eq. (9) relates the face temperature of the shared face between the cells I and I+1 to all the other face temperatures of the two cells, as well as the heat generation rate and the activity levels of the two cells. Equation (9) may be written for each cell face in the floor plan.

At the boundaries of the overall floor plan, convective boundary conditions are imposed. On the x=0 boundary of the floor plan, for example,

$$hA(T_\infty - T_{boundary})\Big|_{x=0} = q_{bf}\Big|_{x=0} \tag{10}$$

where *h* is thermal conductance per unit area and *A* is boundary surface area. Eq. (10) relates the boundary face temperature $T_{boundary}$ to the cell face temperatures of the near-boundary cell and its heat generation rate and activity through Eq. (8).

Equations (7)-(10) form a complete description of all cell, interior face and boundary face temperatures in the floor plan. The corresponding algebraic equation set is solved using a direct solution technique.

3.3 Boundary Conditions for Thermal Models

Our thermal model incorporates buried oxide layer and FinFET device layer. Heat transfer through other layers is modeled as boundary conditions imposed to six faces of circuit block boundary with Eq. (10). For the boundary conditions we made some valid assumptions based on previous work. Most of the heat dissipates through bulk-Si (wafer) and heat sink to air and this heat conduction is limited by the convection flow from heat sink to air. This convection flow has 0.6~0.8 K/W of thermal resistance and the thermal resistance of heat spreader and heat sink is less than 10% of this value [12][13]. For thermal resistance of 0.7 K/W and die size of 1cm^2, we obtain $h_{bottom}=1.4\times10^4$ $W/K/m^2$ as thermal conductance per unit area of the bottom face to wafer and heat sink. For the top face to metal layers and PCB, we assume that the heat dissipation through this is less than 5% of total power dissipation. Ambient temperature is held at 45°C for both top and bottom condition. For lateral heat transfer, we assume that the circuit block is surrounded by silicon oxide and its temperature is not disturbed much by heat generation from other circuit blocks.

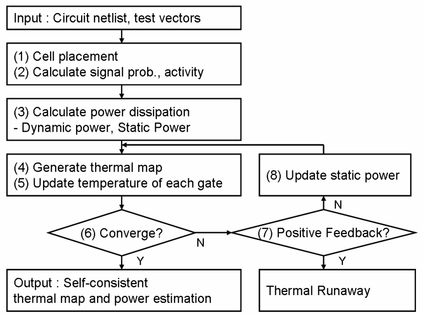

Input : Circuit netlist, test vectors
(1) Cell placement
(2) Calculate signal prob., activity
(3) Calculate power dissipation
- Dynamic power, Static Power
(4) Generate thermal map
(5) Update temperature of each gate
(8) Update static power
N
(6) Converge?
N
(7) Positive Feedback?
Y
Y
Output : Self-consistent
thermal map and power estimation
Thermal Runaway

Figure 3. Flow chart for self-consistent calculation of total power consumption and temperature.

4. SELF-CONSISTENT TEMPERATURE/ POWER ESTIMATION

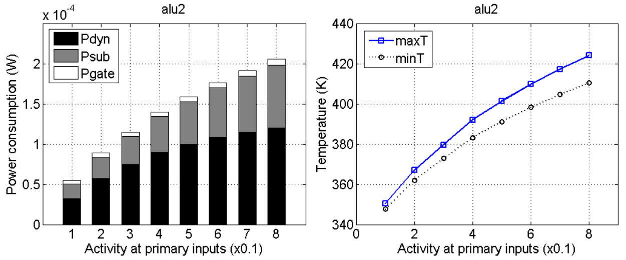
Fig. 3 shows the flow chart for self-consistent temperature and power estimation. In the first step, gate-level placement is performed and the length of each wire is estimated with half perimeter method [14]. This wire length information is used to calculate wire capacitance for dynamic power calculation in step 3. In step 2, with given input patterns, we evaluate signal probability (probability that a node is '1') and switching activity (probability that the value changes) of each node. In step 3, we calculate static and dynamic power dissipation with information obtained so far. In step 4 and 5, gate-level temperature distribution is computed with the proposed thermal models and heat generation values obtained from step 3. Following that, the temperature information of each gate is updated. In step 6, we perform a convergence test: temperature of each gate is compared with the previous value. If it saturates, self-consistent temperature map and power consumption is generated as the final output. If maximum update in temperature during an iteration is higher than that in the previous iteration, thermal runaway occurs. In other cases, static power is recalculated based on current temperature information and thermal map is generated again.

The temperature balancing heat generation (P_{gen}) and heat dissipation (P_{dis}) gives a steady-state solution. P_{gen}=P_{dis} allows only one steady-state solution[1] since P_{dis} is a linear function of temperature and P_{gen} is convex-up with respect to temperature because its temperature dependency is dominated by sub-threshold leakage (Eq. 4). Our algorithm finds proper solution if it exists. If not, the iteration loop fails to converge and it is equivalent to thermal runaway condition.

5. RESULTS AND DISCUSSIONS

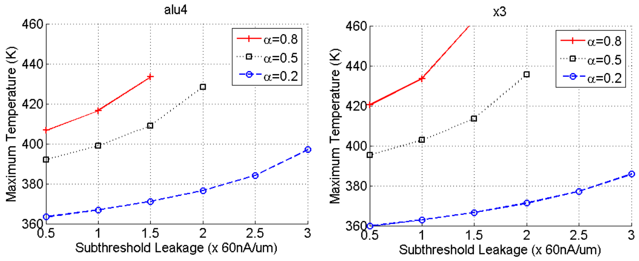
Fig. 4 shows total power consumption and temperature under different PI activity for *alu2* benchmark circuit [15]. V_{dd}=1V, I_{sub}=150nA/μm and *f*=1GHz are assumed for the simulation. It can be seen that the ratio of leakage to dynamic power dissipation

[1] There may be two temperatures that satisfy the equation, but one of them is a meta-stable point.



(a) **(b)**

Figure 4. (a) Power consumption, and (b) temperature under different primary input activity for *alu2*.



(a) *alu4* **(b) *x3***

Figure 5. Sub-threshold leakage current at room temperature vs. maximum temperature.

does not vary significantly with activity: portion of static power component increases from 30% to 38%. High activity inputs result in large active power consumption and increased temperature. The increased temperature (at high activity) increases the sub-threshold leakage, maintaining the ratio of leakage to active power. This emphasizes the need for self-consistently solving temperature and leakage power.

Fig. 5 shows the static temperature rise in *alu4* and *x3* benchmark circuits [15] by varying sub-threshold leakage. Note that the sub-threshold leakage on x-axis is at room temperature and α is the average switching activity of primary inputs. For activity of 0.5, thermal runaway occurs for the sub-threshold leakage of 150nA/μm. It can be seen that *the maximum affordable sub-threshold leakage is dependent on the worst case input activity of the circuit block.* For example, at activity of 0.5 and maximum temperature rise of 430K, the sub-threshold leakage at room temperature should not increase above 120nA/μm for *alu4* and *x3*. This results in an upper bound on acceptable sub-threshold leakage.

The above discussion clearly shows the dependence of thermal runaway on the circuit level parameters such as input activity and transistor level parameters such as sub-threshold leakage (transistor V_t). To reduce the sub-threshold leakage for efficient thermal design, V_t can be increased. However, that will result in lower on-current and hence degraded performance. Our proposed methodology will, therefore, help in optimizing/trading-off transistors/circuits for performance and thermal considerations early in the design phase.

The limit of on sub-threshold leakage can be increased by improving the package quality i.e., heat spreader and heat sink. Fig. 6 quantifies the maximum temperature rise for different values of package thermal resistances. For the thermal resistance of 0.8K/W, I_{sub}=120nA/μm causes thermal runaway. With the package of better heat dissipation 0.6K/W, however, I_{sub} can be increased to 150nA/μm. This clearly emphasizes the need to improve the package quality for high performance operations.

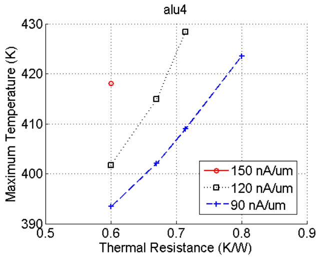


Figure 6. Maximum temperature rise in *alu4* for different values of thermal resistances from device layer to ambient.

6. CONCLUSION

In this work we proposed a methodology to predict thermal runaway by self-consistently solving total power consumption and temperature. We implemented the proposed algorithm and verified it for 28nm FinFET based circuits. FinFETs suffer more from self-heating and less efficient heat dissipation compared to bulk-MOSFETs. The result shows the design trade-off between the input activity of a circuit block, sub-threshold leakage at room temperature and the thermal resistance of the package. We observed that in FinFET circuits, thermal runaway can occur at the ITRS specified sub-threshold leakage (150nA/μm, high-performance) for a nominal activity of 0.5 and typical package thermal resistance.

7. REFERENCES

[1] E. J. Nowak et al., *IEEE Trans. Circ. and Dev. Mag.*, pp. 20–31, Jan-Feb 2004.

[2] L. T. Su et al., *IEEE Trans. Electron Devices*, vol. 41, pp. 69–75, 1994.

[3] E. Pop et al., *IEDM 2003*, pp. 883–886, 2003.

[4] *International Technology Roadmap for Semiconductors*, 2005, available online: http://public.itrs.net.

[5] K. Kanda et al., *IEEE J. of Solid-State Ckts.*, no. 10, pp. 1559–1564, Oct. 2001.

[6] P. Tadayon, *Intel Technology J.*, vol. 3, pp. 1–8, 2000.

[7] K. Roy et al., *Proc. of IEEE*, vol. 91, no. 2, pp. 305–327, Feb. 2003.

[8] S. Mukhopadhyay et al., *IEEE Trans. VLSI System*, pp. 716–730, Aug. 2003.

[9] *Taurus Device Simulator*, Synopsys Inc.

[10] Y. Taur and T. H. Ning, *Fundamentals of Modern VLSI Devices*, Cambridge Univ. Press, 1998, ch. 3, pp. 128.

[11] *Predictive Technology Model*, Arizona State University. http://www.eas.asu.edu/ptm/.

[12] K. Skadron et al., Tech. Rep. CS-2003-08, U. of Virginia Dept. of Computer Science, Apr. 2003.

[13] R. Mahajan et al., *Intel Technology J.*, vol. 6, pp. 62–75, 2002.

[14] M. Sarrafzadeh and C. K. Wong, *An Introduction to VLSI Physical Design*, McGraw-Hill, 1996, ch. 2, pp. 70.

[15] S. Yang, *MCNC'91*, Jan. 1991.