Area Optimization for Leakage Reduction and Thermal Stability in Nanometer Scale Technologies

Ja Chun Ku and Yehea Ismail

EECS Department
Northwestern University
Evanston, IL 60208
{jck273, ismail}@ece.northwestern.edu

Abstract - Traditionally, minimum possible area of a VLSI layout is considered the best for delay and power minimization due to decreased interconnect capacitance. This paper shows however that the use of minimum area does not result in the minimum power and/or delay in nanometer scale technologies due to thermal effects, and in some cases, may result in thermal runaway. A methodology using area as a design parameter to reduce the leakage power, and prevent thermal runaway is presented. A 16-bit adder example in a 70nm technology shows a total power savings of 17% with 15% increase in area, and no increase in delay. The power savings using this technique are expected to increase in future technologies.

I. Introduction

As CMOS devices continue to scale down, the decrease in transistor threshold voltage in order to maintain the performance has resulted in an exponential increase in the subthreshold leakage current [9]. The leakage power that mainly comes from the subthreshold current has already become comparable to the dynamic power in many applications, and it is projected to dominate the total power in sub-100nm technologies, especially at high operating temperatures [10, 11, 17, 20]. Thus, much effort has been put in order to minimize the power consumption through suppressing the subthreshold current [12, 18, 20].

Traditionally, minimum possible area is considered the best for a VLSI layout as it minimized both delay and power consumption due to the decreased interconnect capacitance. However, the use of minimum area has also resulted in an increase in the power density of circuit modules, and hence, the junction temperature increased, which has an exponential impact on the subthreshold current. In other words, minimum area may no longer be the optimum point for power due to the leakage. Furthermore, the use of minimum area may cause thermal runaway for some designs as the leakage power and the temperature are locked into an increasing positive feedback loop. Therefore, controlling power density has to be a crucial part of the design in nanometer scale technologies where leakage is the dominant power dissipation source.

In this paper, area is used as a design parameter to reduce the power density (junction-to-ambient thermal resistance), and it is shown that minimum area does not correspond to the optimum point for power minimization in nanometer scale technologies. By using a larger area for a hot module, its junction temperature is lowered, which reduces the leakage power significantly. However, on the other hand, the increase in interconnect ground capacitance makes the dynamic power increase. The delay traditionally decreases as the temperature drops, but in nanometer scale technologies where the supply voltage is low ($V_{dd} \approx 1$V), the improvement in the delay is very small due to the negative temperature dependence of the threshold voltage. Thus overall, an increase in the area results in a slight increase in the delay. In addition, area can also be used in the design to guarantee thermal stability of a system. This paper provides an analysis of thermal stability condition, and use of the area to prevent thermal runaway.

The next section presents the models used in this work for power and delay as functions of temperature. The thermal model relating power, temperature, thermal resistance, and area is also presented along with the derivation of an analytical expression for steady-state temperature calculation. In the third section, the impact of area optimization is illustrated with a 16-bit adder example in a 70nm technology using the models presented. The fourth section discusses the condition for thermal stability, and shows how the area can be used to prevent thermal runaway using an analytical methodology. Finally, the last section concludes the paper with a summary.

II. Power, Delay, and Thermal Models

In order to evaluate the effect of area on power, delay, and temperature, the first two subsections develop temperature-dependent power and delay models, and compare them to SPICE BSIM3v3 models. Then, the relationship between temperature, power, thermal resistance, and area is modeled in the third subsection with a closed-form expression for steady-state temperature that includes electrothermal coupling.

A. *Power Model*

Power dissipation in CMOS circuits can be divided into two major components

$$P = P_{dynamic} + P_{leakage} \tag{1}$$

There is a third component, short-circuit power that results from a direct-path current when both the pull up and pull down networks are simultaneously on while inputs switch. However, short-circuit power is usually small compared to the other two components of total power, and also is expected to become even less significant as technology scales down [3]. Thus, short-circuit power is ignored to simplify the power models [6, 12].

Dynamic power, $P_{dynamic}$ is given by

$$P_{dynamic} = Na\left(C_g + MCF_pC_c\right)V_{dd}^{\,2}f \tag{2}$$

where N is the number of gates in the design, a is the average switching activity factor, V_{dd} is the supply voltage, and f is the operating frequency. The average load capacitance driven by a gate can be subdivided into the ground capacitance, C_g and the coupling capacitance, C_c. The coupling capacitance is multiplied by MCF_p, the average Miller coupling factor for average power consumption that takes into account the relative switching patterns of the interconnect wires. In nanometer scale CMOS circuits, the coupling capacitance is more dominant compared to the ground capacitance due to the increased interconnect density and smaller feature sizes. To a very good approximation, the capacitance is temperature-independent, and hence, the dynamic component of the power can be assumed to be temperature-independent unless the operating frequency is indirectly affected by the temperature. In this paper, area of a small unit such as a module is considered for optimization, and thus, the chip operating frequency is assumed to

0-7803-9451-8/06/$20.00 ©2006 IEEE.

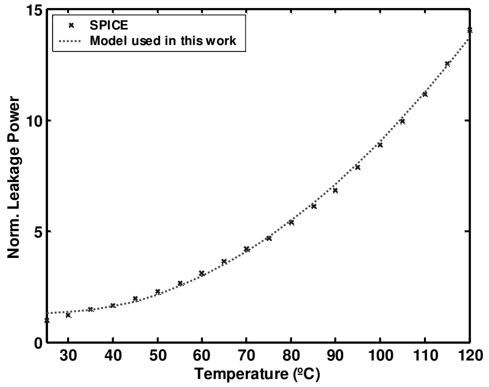

Figure 1. Temperature model of the leakage power compared to SPICE.

be kept constant.

The other major component is the leakage power, $P_{leakage}$, which is dominated by the subthreshold current, and thus it is expressed as

$$P_{leakage} = NI_{sub}(T)V_{dd} \tag{3}$$

The temperature dependence of the average subthreshold current, I_{sub} for a gate can be written as a quadratic function of the temperature as similarly done by Su et al. in [26].

$$I_{sub}(T) = WI_{sub0}\left\{c_1(T_j - T_0)^2 + c_2(T_j - T_0) + c_3\right\} \tag{4}$$

W and I_{sub0} are the average gate width and subthreshold current at nominal temperature, T_0 (which is usually 25ºC), respectively. T_j is the junction temperature, and c_1, c_2, c_3 are constants. The reason that a quadratic function is used rather than an exponential one is because it makes derivation of the expression for steady-state temperature easier in the following subsection. Figure 1 compares the temperature model of the leakage power with a SPICE result in the BPTM [1] 70nm technology. Note that the leakage power more than doubles for every 30ºC rise in the temperature.

The gate oxide leakage current which is ignored in this paper is also on the trend of becoming comparable to the subthreshold current as the thickness of the gate oxide layer shrinks with technology scaling. Unfortunately, SPICE BSIM3v3 (level 49) does not model gate oxide leakage. However, gate oxide leakage is expected to be small compared to the subthreshold leakage in the BPTM 70nm technology used in this paper as can be seen from the data in [16]. Furthermore, gate oxide leakage is relatively temperature-independent, which means that it will be even smaller compared to the subthreshold leakage at high temperatures.

The total power consumption per gate is therefore given by

$$P/N = a(C_g + MCF_pC_c)V_{dd}^2 f + WI_{sub0}\left\{c_1(T_j - T_0)^2 + c_2(T_j - T_0) + c_3\right\}V_{dd} \tag{5}$$

B. Delay Model

As for the delay model, an expression based on the alpha-power law is used [12, 21]

$$D = N_c \frac{(C_g + MCF_dC_c)V_{dd}}{2I_{Di}(T)} \tag{6}$$

where N_c is the number of gates on the critical path, and MCF_d is the average Miller coupling factor for worst-case delay. I_D is the average drain current at saturation region for each gate, and it can be expressed as [9]

$$I_D(T) = KWv_{sat}(T)(V_{dd} - V_t(T))^{\alpha} \tag{7}$$

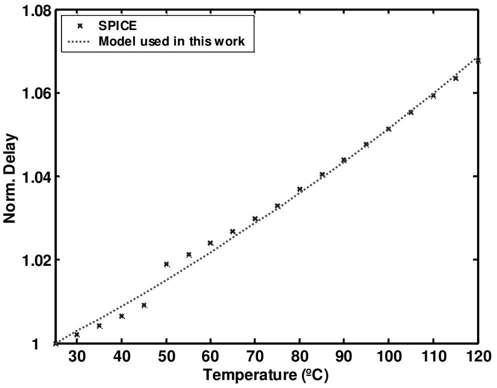

Figure 2. Temperature model of the delay compared to SPICE.

K is a constant specific to a given technology. v_{sat} is the saturation velocity, and α is the velocity saturation index whose value is between 1 and 2 in the deep submicron region [21]. The saturation velocity is given by

$$v_{sat} = \mu E_c \tag{8}$$

where μ and E_c are mobility and the critical electric field for velocity saturation, respectively. Although both saturation velocity and mobility have a negative temperature dependence, its magnitude is weaker for saturation velocity because the value of E_c also becomes higher as the temperature is raised. Note that the drain current stays in the saturation region for almost the entire duration of the transition. Thus, the temperature dependence of the drain current follows that of the saturation velocity rather than the mobility. The temperature dependence of saturation velocity is almost linear [7, 19], and can be written as

$$v_{sat}(T) = v_{sat}(T_0) - \eta(T - T_0) \tag{9}$$

η is the saturation velocity temperature coefficient whose value obtained from SPICE simulations is around 140 ms^{-1}/ºC for a 70nm technology. The threshold voltage also decreases as the temperature is raised, and is given by

$$V_t(T) = V_t(T_0) - \kappa(T - T_0) \tag{10}$$

κ is the threshold voltage temperature coefficient whose value is about 0.7mV/ºC in deep submicron technologies [27]. When the supply voltage is high, the effect of the saturation velocity dominates the overall temperature dependence of the drain current since the change in $(V_{dd}–V_t(T))^{\alpha}$ is relatively insignificant. However, when the supply voltage is low (close to 1V), the change in $(V_{dd}–V_t(T))^{\alpha}$ becomes more important, and cancels the temperature dependence of the drain current due to the saturation velocity. Figure 2 compares the temperature model of delay with a SPICE result in BPTM [1] 70nm technology operating at V_{dd} = 1.1V. The delay increases only by 6-7% after a temperature rise of 80-90ºC.

C. Thermal Model

The heat generated from a chip is dissipated through the silicon substrate, and the cooling system in the package. The heat flow in the package is a function of many parameters such as geometry, flux source and placement, package orientation, next-level package attachment, heat sink efficiency, and method of chip connection [15]. In this paper, we consider a typical flip-chip C4 package adapted from models by Kromann in [15] as shown in Figure 3. Most of the heat is dissipated through the heat sink that is usually attached to the back-side of the silicon substrate. This constitutes the primary heat transfer path where the heat generated is conducted upwards through the silicon to the thermal paste,

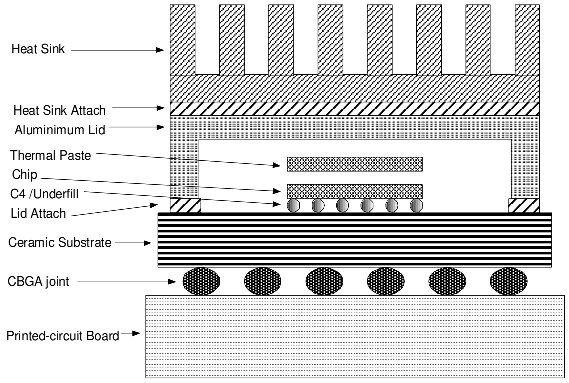

Figure 3. A typical flip-chip C4 package.

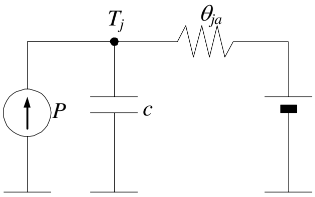

Figure 4. One-dimensional chip thermal model.

aluminum cap, heat sink attach, and heat sink, then convectively removed to the ambient air.

Since different parts of a chip have different activities and power densities, there is a variation in the thermal profile across a chip [4]. Accurate thermal modeling of a whole chip would thus require a 3-D analysis. However, the target of the area optimization technique proposed in this paper is not the entire chip, but only a small unit of a chip such as a module that is hot and leaking much (hot spot), and there is not much variation in temperature within such units. Hence, the following 1-D model is a valid approximation for its thermal analysis.

$$\theta_{ja} c \frac{dT_j}{dt} + T_j = P(T_j)\theta_{ja} + T_a \tag{11}$$

where θ_{ja} is the junction-to-ambient thermal resistance of the silicon substrate and the package, c is the thermal capacitance of the system, T_j is the junction temperature, P is the power dissipation, and T_a is the ambient air temperature which is usually assumed to be 45ºC. Figure 4 shows an equivalent electrical circuit for the thermal model [14]. Note that power is a function of the junction temperature with a positive dependence as it can be seen in (5). A rise in the temperature results in an increase in the power, which in turn, raises the temperature even higher, thus creating a positive feedback loop until the system reaches a steady-state (electrothermal coupling).

The thermal resistances of the silicon, the aluminum cap and the heat sink attach are small, and their contribution to the temperature drop can be omitted for a first-order analysis [15]. Hence, the junction-to-ambient thermal resistance can be expressed as

$$\theta_{ja} = \theta_{thermalpaste} + \theta_{heat\sin k} \tag{12}$$

It is shown by Kromann in [15] that the thermal paste resistance is reduced as the chip area increases. This is because the thermal resistance can be written as [5]

$$\theta = {R_{th}}\Big/{A_c} \tag{13}$$

where R_{th} is the unit thermal resistance, and A_c is the cross-sectional area. An increase in the chip area directly increases the area of the thermal paste placed above it, thus assuming the chip area equals to the thermal paste area, the steady-state temperature (when the time derivative of the junction temperature equals to zero) can be written as

$$T_j = \left({P(T_j)}\Big/{A_{chip}} \right) R_{thermalpaste} + P(T_j)\theta_{heat\sin k} + T_a \tag{14}$$

where $P(T_j)/A_{chip}$ represents the power density of the chip, and $R_{thermalpaste}$ is the unit thermal resistance of the thermal paste. Convective thermal resistance of the heat sink, $\theta_{heatsink}$ is affected less by the chip area since the heat is usually spread out more uniformly (using a heat spreader) before it reaches the heat sink. However, in case of adapting an advanced fan heat sink as it is commonly done in today's technology, the heat sink resistance becomes small enough that the thermal paste resistance takes up the majority of the total junction-to-ambient thermal resistance (more than 60%) [15]. Therefore, increasing an area in the chip can significantly lower the junction temperature.

In order to derive an expression for the steady-state temperature after the electrothermal coupling, the temperature dependence of the power in (5) is substituted in (11). After some rearrangement, the following equation can be obtained.

$$\begin{aligned}
&\theta_{ja} c \frac{dT_j}{dt} = AT_j^{\,2} + BT_j + C \\
&where \\
&A = c_1 \theta_{ja} NWI_{sub0} V_{dd} \\
&B = \theta_{ja} NWI_{sub0} V_{dd} (c_2 - 2c_1 T_0) - 1 \\
&C = T_a + \theta_{ja} NV_{dd} \left\{ a(C_g + MCF_p C_c) V_{dd} f + WI_{sub0} \left(c\, T_0^{\,2} - c_2 T_0 + c_3 \right) \right\}
\end{aligned} \tag{15}$$

For steady-state, the time derivative of the junction temperature is equal to zero. Then, the quadratic equation on the right-hand side in (15) is solved to obtain the steady-state temperature value.

$$T_j = \frac{-B \pm \sqrt{B^2 - 4AC}}{2A} \tag{16}$$

Since the equation is quadratic, there are two solutions. The smaller one of the two corresponds to a stable point while the other corresponds to a metastable point. The stability of the two solutions is illustrated in Figure 5 which plots the time derivative of temperature, T_j' against the temperature, T_j. Positive T_j' indicates that T_j is increasing while a negative T_j' indicates that T_j is decreasing. Hence, if the curve crosses $T_j' = 0$ with a negative slope, a small perturbation to either direction would bring T_j back to the point where $T_j' = 0$. On the other hand, if the curve crosses $T_j' = 0$ with a positive slope, any small perturbation would make T_j run

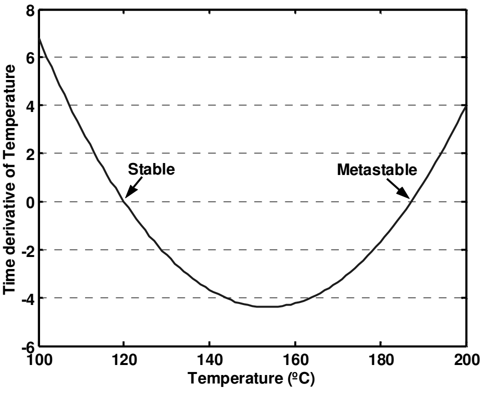

Figure 5. Solutions for the steady-state temperature.

away from the point. Thus, the relevant solution for steady-state temperature is only the one that corresponds to a stable point. After calculating the steady-state temperature, the steady-state value for the power consumption can be easily obtained as well by using (5).

III. Area Optimization for Leakage Reduction

In this paper, when an area increase of a design is considered, the length of the interconnect wires and the space between the wires are increased by the same ratio, while the height, thickness, and width of the wires are kept unchanged. Increasing the gate size along with the layout area was also initially considered in the optimization process. However, it was found that the increase in the load capacitance (which affect both delay and dynamic power) and the leakage power by increasing the gate size makes the initial gate size the optimum in terms of power-delay product. This is true also because of the fact that the interconnect capacitance does not change significantly with area due to the cancellation between ground and coupling components in the capacitance (which will be further explained in the following paragraphs). Thus, the sizes of the devices are kept unchanged.

The area scaling factor is called *x* in this paper. Thus, both the length of the wires and the space between the wires are increased by a factor of $x^{1/2}$. Figure 6 shows the effects of an area scaling by a factor if *x* on the interconnect wires. Increasing the area has two major impacts on the power and delay. The first impact is due to a decrease in the thermal resistance as shown in (13). The decrease in the thermal resistance results in a lower steady-state temperature as explained in the previous section. The subthreshold current will decrease exponentially, which in turn, results in a lower junction temperature. This means that the steady-state value for the power consumption becomes significantly lower. The drain current from (9) will also increase as the temperature drops, resulting in a slightly better delay. However, this improvement in the delay is not significant as discussed earlier.

The second impact is caused by changes in the load capacitances as the interconnect wires become longer. The coupling capacitance of the wires increases linearly as the length of the wires increases by $x^{1/2}$. However at the same time, the coupling capacitance of the wires also decreases superlinearly with an exponent around 1.34 as the space between the wires widens according to the work by Sakurai in [22]. Since the space between the wires is increased by $x^{1/2}$, the coupling capacitance decreases by $(x^{1/2})^{1.34}$. This superlinear decrease in the coupling capacitance outweighs the linear increase due to increasing wire length, resulting in an overall decrease in the coupling capacitance by $x^{0.17}$. On the other hand, the ground capacitance increases by $x^{1/2}$ as the length of the wires increases, countering the effects of the decreasing coupling capacitance. Hence, the total interconnect capacitance after an area scaling by *x* can be described by

$$x^{0.5}C_g + MCFx^{-0.17}C_c \qquad (17)$$

where C_g and C_c are the ground and coupling capacitance of the wires before area scaling, respectively. The total interconnect capacitance may initially decrease slightly for small *x* since coupling capacitance is more dominant compared to ground capacitance in nanometer scale technologies. However, the total interconnect capacitance will start to increase for larger *x* because the magnitude of the positive exponent of the ground capacitance is greater than that of the negative exponent of the coupling capacitance. The increase in the interconnect capacitance results in an increase in both power and delay. The observations above indicate that there is an optimum area that designers can choose to reduce leakage.

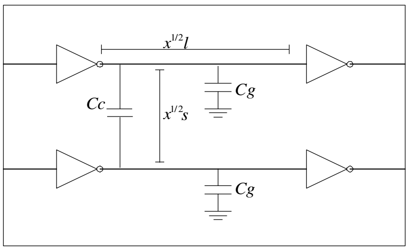

Figure 6. Changes in the wire dimensions after an area scaling by a factor of *x*.

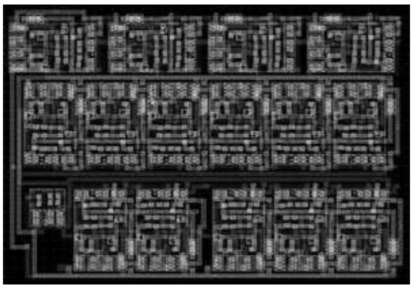

Figure 7. Layout of the 16-bit adder.

In order to evaluate the effectiveness of the area optimization technique proposed in this paper, a 16-bit adder is laid out for 70mn BPTM [1] technology operating at V_{dd} = 1.1V (Figure 7). Then, the models presented in the previous section are used to simulate the impact of area scaling on junction temperature, power, and delay. The value of the junction-to-ambient thermal resistance, which depends on the packaging technology, was chosen such that the junction temperature is around 120ºC. Figure 8 illustrates the relationship between the junction temperature and the area. Notice that the temperature initially drops significantly when the area increases, then starts to saturate as *x* becomes larger. The initial big drop in the temperature is due to the exponential reduction in the leakage power. However, as more leakage power is eliminated, the dynamic power becomes more dominant, and increases slightly for larger *x* due to the increase in the interconnect capacitance, countering the decrease in the leakage power. This trend implies that the effect of the temperature drop on power and delay will be strongest during the initial increase in the area, and start to decay for larger *x*. Figure 9 shows how power and delay are affected as the area is increased. As expected, there is a big drop in power in the beginning, then it soon starts to saturate, following the behavior of the temperature. It can be seen from Figure 8 that the delay improves slightly in the beginning due to the initial decrease in the interconnect capacitance as discussed above (plus the temperature drop), however it soon starts to increase as *x* becomes larger. The reason the delay keeps increasing without saturating is that the

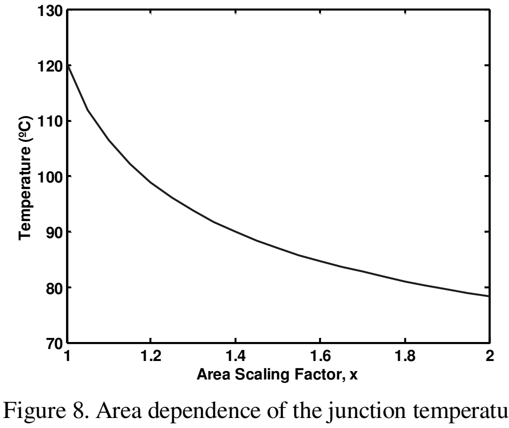

Figure 8. Area dependence of the junction temperature.

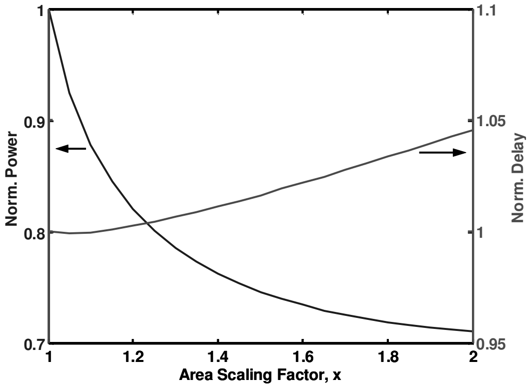

Figure 9. Area dependence of the power and delay.

temperature effect is not significant enough to counter the increase in the interconnect capacitance. However, the increase in the delay due to an area increase is still negligible (*increases only about 4-5% even when the area is doubled*). Note that the power consumption is reduced about 17% when the area is increased by 15% with almost no change in the delay.

The choice of the area would differ between each design depending on the design goals and constraints. It also strongly depends on how much weight the area cost carries. Although it depends on the purpose of the chip, area cost is becoming less significant as technology scales down especially in the case of high-performance ICs. In future technologies, delay will have a higher sensitivity to area due to smaller supply voltage applied. However at the same time, leakage power will dominate the total power even more, and since the leakage power increases exponentially with technology scaling, the reduction in the power through an area increase will easily exceed the increase in the delay. Furthermore, coupling capacitance will take up a larger fraction of the total interconnect capacitance, which will counter the effect of increased ground capacitance. Therefore, the proposed area optimization technique will prove to be even more useful in future technologies.

IV. Area Optimization for Thermal Stability

The possibility of thermal runaway in current and future technologies has been discussed in some literatures [14, 24]. The main cause for thermal runaway is the rapid increase in leakage current as the technology scales down. A large leakage power drives a high junction temperature as described by (12), and the new temperature pushes the leakage power up to an even higher value creating a positive feedback loop. If the gain of this loop is higher than a certain value, the feedback results in thermal runaway, and the system fails.

Traditionally, solutions to the thermal management problems have been mostly focused on the package cooling technology. However, there is usually a disparity between the maximum power and typical power consumed, and this gap is increasing with technology scaling. The package designed for worst-case (maximum power) is too expensive since the cooling cost increases nonlinearly with thermal dissipation [13]. Thus, packages today are typically designed for the *typical* power consumption to significantly reduce the package cost, and other dynamic alternatives such as frequency scaling, voltage scaling, and global clock gating are used in order to maintain the temperature in the permissible range when the power consumption increases beyond the typical case [13, 24]. In nanometer scale technologies where leakage dominates the power consumption, an increase in the power will not just raise the temperature, but also may result in thermal runaway due to the electrothermal coupling effect.

The stability condition can actually be derived from (16). The term inside the square root has to be positive or zero for real solutions to exist for the temperature. A negative value inside the square root corresponds to thermal runaway since that means there is no real stable solution. Therefore, the following condition has to be satisfied for the system to be stable.

$$B^2 - 4AC \geq 0 \tag{18}$$

It can be seen from (18) that parameters such as the supply voltage, operating frequency, junction-to-ambient thermal resistsance, and ambient temperature can cause thermal runaway when their values increase above certain critical values. In other words, such parameters can also be used to control and prevent thermal runaway.

As mentioned above, there are existing dynamic techniques that prevent thermal runaway. Typical examples are frequency scaling, voltage scaling, and global clock gating [24]. When the power of a system exceeds a certain threshold value (that may cause failure of the system) the operating frequency and/or the supply voltage of the chip are immediately dropped to a level where the power would be reduced enough to maintain the stability of the system. As for global clock gating, the global clock is stopped until the temperature falls down to a safe value. The common drawback of these techniques is that the whole chip has to be slowed down even though a hot small area (hot spot) may be the only cause of thermal runaway, which is often the case. In addition, all these techniques require sensors on chip for real-time temperature sensing which are sensitive to lithographic variations and supply-current variations causing imprecision [24].

An alternative way to prevent thermal runaway proposed in this paper is to increase the local area of hot spots (reducing its junction-to-ambient thermal resistance). This way, the performance of the chip is likely to be unaffected since there is no change in the delay for a small increase in area as it was shown in the previous sections. Hence, the operating frequency of the chip does not need to be dropped. From (18), the critical value of the junction-to-ambient thermal resistance, θ_c for the system to be stable can be derived by rearranging (18) as a quadratic function of θ_c.

$$\begin{aligned}
&X\theta_c^{\,2} + Y\theta_c + Z = 0 \\
&\textit{where} \\
&X = N^2 W I_{sub0} V_{dd}^{\,2}\left\{W I_{sub0}\left(c_2^{\,2} - 3c_1^{\,2}T_0^{\,2} - 4c_1c_3\right) - 4c_1 a\left(C_g + MCF_pC_c\right)V_{dd}f\right\} \\
&Y = -NWI_{sub0}V_{dd}\left(2c_2 - 4c_1T_0 + 4c_1T_a\right) \\
&Z = 1
\end{aligned} \tag{19}$$

The solution of the quadratic equation in (19) is

$$\theta_c = \frac{-Y - \sqrt{Y^2 - 4X}}{2X} \tag{20}$$

As it can be seen from (20), there is only one solution for (19) because the other solution leads to a lower junction temperature as the thermal resistance increases, which physically does not make sense. Thus, the value of junction-to-ambient thermal resistance has to be equal or smaller than θ_c in order to maintain thermal stability of the system. A junction-to-ambient thermal resistance that is greater than θ_c leads to thermal runaway. Figure 10 shows the transient response of temperature for two different values of thermal resistance. The minimum area, A_{min} that is required for thermal stability can be therefore derived as well by relating (20) to (12) and (13).

$$A_{\min} = \frac{R_{thermalpaste}}{\theta_c - \theta_{heat\sin k}} \tag{21}$$

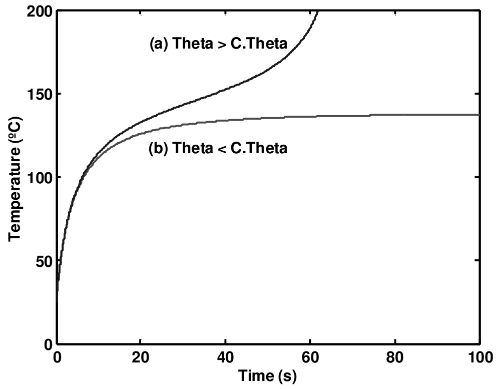

Figure 10. Transient response of temperature for (a) theta > critical theta (b) theta < critical theta.

V. Conclusion

It was shown in this paper that area can be used as a degree of freedom to reduce leakage, and prevent thermal runaway in nanometer scale technologies. Increasing the area reduces the thermal resistance, thus lowering the junction temperature. The power is initially reduced significantly as the area is increased because the drop in the temperature results in an exponential decrease in the leakage power. However, the reduction in the power saturates as the area is increased further since the dynamic power rises slowly, and the leakage power becomes smaller. Delay worsens slightly as the area is increased due to the increasing interconnect capacitance. However, the change in the delay is relatively insignificant. An analytical model was presented for the calculation of steady-state junction temperature after the electrothermal coupling. The area optimization technique for leakage reduction was evaluated using the power, delay, and thermal models presented with on a 16-bit adder example in a 70nm technology. It was shown that by increasing the area of a hot spot a significant power reduction is obtained. This trend is expected to continue as technology scales down because the decrease in the leakage power will easily exceed the increase in the delay. The possibility and condition for thermal runaway were also discussed. It was shown that the method of increasing a local area where it may cause thermal runaway can be used to maintain thermal stability without slowing down the whole chip. An analytical design methodology has been used to derive the required area for thermal stability.

References

[1] http://www-devices.eecs/berkeley.edu/~ptm/mosfet.html

[2] K. Banerjee, S-C. Lin, A. Keshavarzi, S. Narendra and V. De, “A self-consistent junction temperature estimation methodology for nanometer scale ICs with implications for performance and thermal management,” in Proc. IEDM, pp. 887-890, 2003

[3] K. Banergee and A. Mehrotra, “A power-optimal repeater insertion methodology for global interconnects in nanometer designs,” in IEEE Trans. Electron Devices, vol. 49, pp. 2001-2007, Nov. 2002

[4] S. Borkar et al, “Parameter variations and impact on circuits and microarchitecture,” in Proc. DAC, pp. 338-342, 2003

[5] A. Chapman, Fundamentals of heat transfer. Macmillan Press, 1987

[6] A. Chatterjee, M. Nandakumar, and I. Chen, “An investigation of the impact of technology scaling on power wasted as short current in low voltage CMOS,” in Proc. ISLPED, pp. 145-150, Aug. 1996

[7] Y. Cheng, K. Imai, M. Jeng, Z. Liu, K. Chen, and C. Hu, “Modeling temperature effects of quarter micrometer MOSFETs in BSIM3v3 for circuit simulation,” in Semicond. Sci. Tech., pp. 1349-1354, 1997

[8] J. Daga, E. Ottaviano, D. Auvergne, “Temperature effect on delay for low voltage applications,” in Proc. DATE, pp. 680-685, 1998

[9] V. De and S. Borkar, “Technology and design challenges for low power and high performance,” in Proc. ISLPED, pp. 163-168, 1999

[10] V. De and S. Borkar, “Low power and high performance design challenges in future technologies,” in Proc. GLSVLSI, pp. 1-6, 2000

[11] D. Genossar and Y. Shemir, “Intel Pentium M processor power estimation, budgeting, optimization, and validation,” in Intel Tech. J., vol. 7, May 2003

[12] R. Gonzalez, B. Gordon, M. Horowitz, “Supply and threshold voltage scaling for low power CMOS,” in IEEE J. Solid-State Circuits, vol. 32, pp.1210-1216, Aug. 1997

[13] S. Gunther, F. Binns, D. Carmean, and J. Hall, “Managing the impact of increasing microprocessor power consumption,” in Intel Tech. J. Q1, 2001

[14] K. Kanda, K. Nose, H. Kawaguchi, and T. Sakurai, “Design impact of positive temperature dependence on drain current in sub-1-V CMOS VLSIs,” in IEEE J. Solid-State Circuits, vol. 36, pp1559-1564, Oct. 2001

[15] G. Kromann, “Thermal modeling and experimental characterization of the C4/surface-mount-array interconnect technology,” in IEEE Trans. Component, Packaging, and Manufacturing Technology, vol. 18, no. 1, pp. 87-93, Mar. 1995

[16] D. Lee, D. Blaauw, and D. Sylvester, “Gate oxide leakage current analysis and reduction for VLSI circuits,” in IEEE Trans. VLSI, vol. 12, pp. 155-166, Feb. 2004

[17] S. Lin, A. Basu, A. Keshavarzi, V. De, A. Mehrotra, K. Banerjee, “Impact of off-state leakage current on electromigration design rules for nanometer scale CMOS technologies,” in Proc. IRPS, pp. 74-78, 2004

[18] K. Nose and T. Sakurai, “Optimization of V_{dd} and V_{th} for low-power and high-performance applications,” in Proc. ASP-DAC, pp. 469-474, 2000

[19] R. Quay, C. Moglestue, V. Palankovski, S. Selberherr, “A temperature dependent model for the saturation velocity in semiconductor materials,” in Material Science in Semiconductor Processing 3, pp. 149-155, 2000

[20] K. Roy, S. Mukhopadhyay, H. Mahmoodi-Meimand, “Leakage current mechanisms and leakage reduction techniques in deep-submicrometer CMOS circuits,” in Proc. IEEE, vol. 91, pp. 305-327, Feb. 2003

[21] T. Sakurai and A. Newton, “Alpha-power law MOSFET model and its application to CMOS inverter delay and other formulas,” in IEEE J. Solid-State Circuits, vol. 25, pp. 584-593, Apr. 1990

[22] T. Sakurai, “Closed-form expression for interconnection delay, coupling, and crosstalk in VLSI’s,” in IEEE Trans. Electron Devices, vol. 40, no. 1, pp. 118-124, Jan. 1993

[23] O. Semenov, A. Vassighi, M. Sachdev, A. Keshavarzi, and C. Hawkins, “Effect of CMOS technology scaling on thermal management during burn-in,” in IEEE Trans. Semiconductor Manufacturing, vol. 16, no. 4, pp. 686-695, Nov. 2003

[24] K. Skadron, M. Stan, W. Huang, S. Velusamy, K. Sankaranarayanan, and D. Tarjan, “Temperature-aware microarchitecture,” in Proc. ISCA, 2003

[25] S. Strogatz, Nonlinear dynamics and chaos. Westview Press, 1994

[26] H. Su, F. Liu, A. Devgan, E. Acar, S. Nassif, “Full chip leakage estimation considering power supply and temperature variations,” in Proc. ISLPED, pp. 78-83, 2003

[27] Y. Taur and T. Ning, Fundamentals of modern VLSI devices. Cambridge University Press, 1998