Compact Thermal Modeling for Temperature-Aware Design

ABSTRACT

Thermal design in sub-100nm technologies will become one of the major challenges to the CAD community. Using temperature as a guideline for design is very important at sub-100nm. In this paper, we first introduce the idea of *temperature-aware* design. We then propose a compact thermal model which can be integrated into modern CAD tools to achieve a temperature-aware design. Finally, we use the compact thermal model in a microprocessor design case study to show the importance of using temperature as a guideline for the design. Results from our thermal model show that temperature-aware design approach can provide more accurate design estimations, and therefore better design decisions and faster design convergence.

1. INTRODUCTION

As CMOS technology is scaled into the sub-100nm region, power density of microelectronic designs increases steadily. For example, average power density of high-performance microprocessors have already reached (50W/cm^2) at 100nm technology, and will soon reach (100W/cm^2) at technologies below 50nm[1]. As a result, average temperature of the die also increases rapidly. Furthermore, local hot spots on the die usually have significantly higher power density than the average, making the local die temperature even higher.

High temperature has a number of negative impacts on microelectronic designs—first, transistor speed is slower at higher temperature because of the degradation of carrier mobility. A back-of-the-envelop calculation shows that a single inverter is about 35% slower at 110°C than at 60°C; Second, temperature dependance of leakage power is significant. Leakage power can be orders of magnitude greater at higher temperatures[2]. To make things even worse, leakage power is exceeding switching power and becoming a dominant source of power consumption in sub-100nm designs[1]; Third, interconnect metal resistivity is also dependent on temperature. For example, resistivity of copper increases by 39% from 1.72$\mu\Omega$-cm at 20°C to 2.39$\mu\Omega$-cm at 120°C. Higher resistivity causes longer interconnect RC delay, and hence performance degradation; Last, but not the least, temperature is strongly related to design reliability. The impact of temperature on reliability can be modeled by Arrenhius Equation $MTF{=}MTF_0\ exp(E_a/k_bT)$, where MTF_0 is mean time to failure at a specified reference temperature, E_a is the activation energy of the failure, k_b is the Boltzmann constant. A well-known example of microelectronics reliability problems is interconnect electromigration. It is obvious from Arrenhius Equation that increasing the temperature will exponentially decrease the mean time to failure, and hence the life time. In a sentence—for future designs, higher operating temperature will have significant negative impacts on performance, power consumption, and reliability.

Based on the above facts, thermal design will become one of the major challenges for the CAD community in sub-100nm designs such as microprocessors, ASICs or System-on-a-Chip (SoC). Existing design methodologies usually simply use worst-case or room temperature when needed. This will lead to significant design estimation errors and hence wrong design decisions and longer design convergence time, as will be seen in a case study in Section 5. Therefore, it is crucial to find a way to properly address the temperature-related aspects of the design flow, and use temperature as a guideline for design.

This paper is organized as follows. Section 2 introduces the idea of *temperature-aware* design. Section 3 proposes a compact thermal model that can be integrated into CAD tools to achieve a temperature-aware design flow. Validation of the model is presented in Section 4. In Section 5, a microprocessor design case study using the compact thermal model shows the importance of using temperature as a guideline for design. Section 6 concludes the paper.

2. TEMPERATURE-AWARE DESIGN

In sub-100nm technologies, early accurate design estimation is key to high-level design convergence and should ensure careful consideration of deep submicron effects (including power, performance, reliability, etc.)[3]. Temperature plays an important role in early, accurate power, performance and reliability estimations. In addition, temperature is also closely related to placement and routing, because, intuitively, a hotter block could be surrounded by several colder blocks. Due to lateral heat transfer among blocks, the hotter block would have a lower temperature than if it were placed near some other hot blocks. Thus, if all other conditions remained the same, this hotter block would consume less power, have shorter delay and longer life time. This means that at sub-100nm, temperature should be included in the cost function in order to achieve optimal placement and routing. Temperature can also affect manufacturability in terms of packaging design and choices of process if the design is thermally limited by a particular process. Figure 1 shows a typical ASIC design flow. We can see from this example that most design stages are temperature-related. Also, we can see that temperature profiles are needed at both functional-block level and standard-cell level during the ASIC design flow. Similar argument also applies to microprocessor and SoC designs.

From above, we see that it is very important to be able to estimate temperature at different granularities and at different design stages, especially the early ones. The estimated temperature can then be used throughout the design flow to perform power, performance, and reliability analyses, together with placement, packaging design, etc. As a result of this, all the decisions made during design use temperature as a guideline and the design is guaranteed to be thermally optimized and free from thermal limitations. We call this type of design methodology *temperature-aware* design. The idea of temperature-aware design is unique because operating

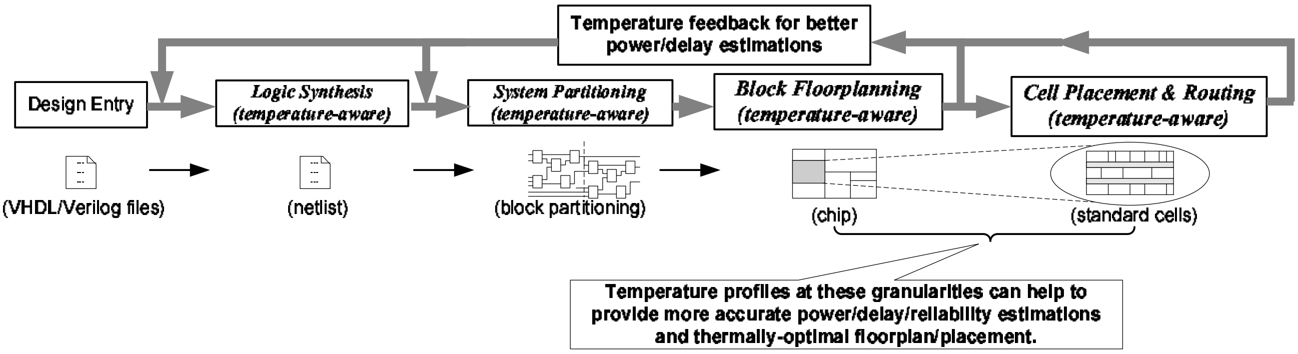
Temperature feedback for better power/delay estimations
Design Entry
Logic Synthesis (temperature-aware)
System Partitioning (temperature-aware)
Block Floorplanning (temperature-aware)
Cell Placement & Routing (temperature-aware)
(VHDL/Verilog files)
(netlist)
(block partitioning)
(chip)
(standard cells)
Temperature profiles at these granularities can help to provide more accurate power/delay/reliability estimations and thermally-optimal floorplan/placement.

Figure 1: An example of temperature-aware ASIC design flow.

temperature is properly considered during the *entire* design flow instead of being considered only at the end of the design flow. There has been some previous work about temperature-related design—for example, in [4], the authors present a design flow from digital simulations to a thermal map at the end of the design. Being able to obtain a post-layout chip-level thermal map is useful, but this design flow cannot be termed as temperature-aware design because none of the intermediate design stages have closely considered temperature-related issues such as power or performance estimations, placement thermal analysis, etc. Thus the design decisions of these stages are not optimized, and the design has to restarts from the beginning if it turns out to be thermally limited.

3. A COMPACT THERMAL MODEL

In order to achieve temperature-ware design, we need a thermal model to estimate operating temperature. Figure 2 shows in detail how a thermal model helps to close the loop for accurate power, performance and reliability estimations. For example, the power model first provides estimated power to the thermal model. The thermal model in turn provides estimated temperature to the power model, and so on. After a few iterations, both power and temperature estimations converge. At that point, temperature-aware power estimation is achieved. Similarly, temperature-aware performance and reliability estimations can also be achieved.

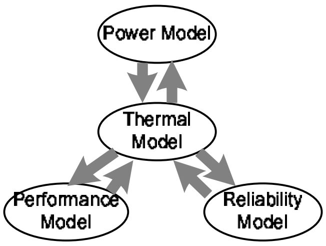

Figure 2: Interactions among thermal model and power, performance and reliability models.

There are a number of existing thermal models for different parts of a microelectronic design. For example, [5] presents a dynamic thermal model only at the microarchitecture level. [6] presents a chip-level thermal model based on full-chip layout. In [7], the authors present package thermal models. In [8], the authors present a thermal modeling approach based on analytical solutions of heat transfer equations, and the model is mainly focused at device level. None of these thermal models have the flexibility to model temperature at arbitrary granularity. Some of them are also computationally intensive. Thus, they are not completely suitable for temperature-aware design. To fulfill the requirements of a temperature-aware design, the thermal model has to be able to provide temperatures at different granularities (circuit structures, standard cells, functional unit blocks, etc.), and at different levels (silicon surface, interconnect, package, etc.). The model also needs to be computationally efficient to avoid time-consuming calculations during high-level, prior-layout design stages. In some cases, the model also should be able to model transient temperature changes. Of course, the model should be reasonably accurate to provide useful temperature estimates.

Here, we propose a *compact* thermal model that meets all the above requirements and can be used to achieve temperature-aware design. Before moving on to modeling details, it is useful to notice that this compact thermal model is a general model and therefore can be applied to different contexts. For example, dynamic thermal management (DTM) is an active research area in computer architecture community[9]. In this context, although at run time, only thermal sensors and DTM methods are needed in order to implement DTM techniques, the thermal model is still very attractive to computer architects, because it helps to simulate and explore different DTM methods. As another example, in the design automation context, we have already argued in Section 1 and 2 that a thermal model is needed to guide CAD, choices in process, circuit style, packaging, etc. People from CAD community and industry may consider the thermal model attractive. In this paper, we are interested in the latter context.

3.1 Model Overview

There is a well-known duality between heat transfer and electrical phenomena. In this duality, heat flow that passes through a thermal resistance is analogous to electrical current; Temperature difference is analogous to voltage. Similar to an electrical capacitor that accumulates electrical charges, thermal capacitance defines the capability of a structure to absorb heat. The rationale behind this duality is that electrical current and heat flow can be described by exactly the same differential equations, if electrical inductance is not considered. The compact thermal model we propose is essentially a thermal RC circuit. Each node in the circuit corresponds to a block at the interested level of granularity. Solving the thermal RC circuit gives the temperatures of each node.

Figure 3(a) shows a modern single-chip CBGA package[10]. Heat generated from the active silicon device layer may be conducted through silicon die to interface material, heat spreader and heat sink, then convectively removed to the ambient air. In addition to this primary heat transfer path, a secondary heat flow path exists from conduction through the interconnect layer, I/O pads, ceramic substrate, leads/balls to the printed-circuit board. Our compact thermal model models all these layers in both heat flow paths, with special emphasis on the primary heat flow path and the on-chip interconnect layer. This is because, detailed temperature profile of these parts is very important for temperature-aware design. In the model, we also consider lateral heat flow within each layer to achieve greater accuracy of temperature estimation. Figure 3(b) shows the thermal RC circuit structure that corresponds to Figure 3(a). Next, we will present the modeling details of each layers along both heat flow paths.

3.2 Primary Heat Flow Path

Figure 4(a) shows an example thermal circuit of a silicon die with only three microarchitecture blocks from [5]. We extend their thermal model for the primary heat flow path by making the model

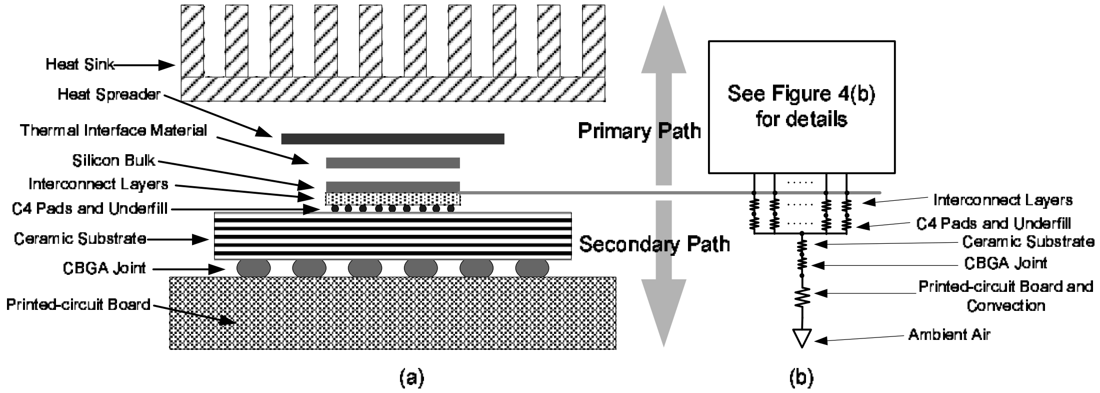

Figure 3: (a) A typical flip-chip, CBGA package with heat sink (adapted from [10]). (b) Corresponding thermal circuit in our thermal model. Thermal capacitors connecting each node to ambient are not shown for clarity.

grid-like and be able to model temperatures at any arbitrary granularity. Figure 4(b) shows our modeling approach with 3x3 grids. Each silicon grid can be of arbitrary aspect ratio and size, which are determined by the interested level of granularity. We add to the model, a layer of thermal interface material that is absent in [5]. In addition, part of the heat spreader that is right under the interface material and the interface material itself is divided into the same number of grids as the silicon die. Other parts in the primary heat flow path are modeled in a similar way as in [5]—remaining part of the heat spreader is divided into four trapezoidal blocks. Heat sink is divided into five blocks: one corresponding to the area right under the heat spreader; four trapezoids for the periphery. Each grid or block corresponds to a node in the thermal circuit. There are vertical and lateral thermal resistors connecting the nodes. Package-to-air thermal resistor is calculated from specific heat-sink configurations and ambient conditions.[1] Each node also has a thermal capacitor connected to the ambience. Power generated from each silicon grid is modeled as a "current source" connected to the corresponding node.

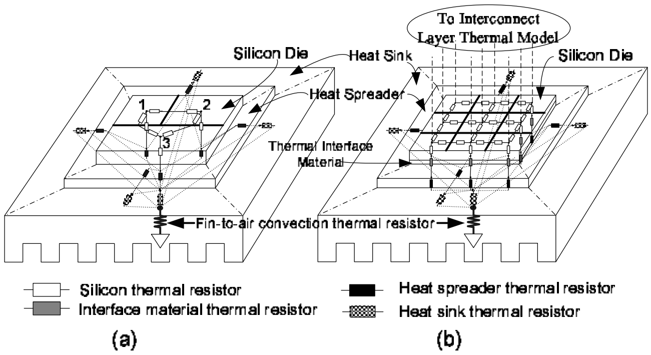
To Interconnect Layer Thermal Model
Silicon Die
Heat Sink
Silicon Die
Heat Spreader
1
2
3
Thermal Interface Material
Fin-to-air convection thermal resistor
Silicon thermal resistor
Interface material thermal resistor
Heat spreader thermal resistor
Heat sink thermal resistor
(a)
(b)

Figure 4: (a) Thermal circuit of a silicon die with 3 microarchitecture blocks, adapted from [5]. (b) Thermal circuit of a silicon die with 3x3 grids, together with thermal interface material, heat spreader and heat sink. (Thermal capacitors and heat sources are not shown for clarity.)

The derivation is mainly based on the fact that vertical thermal resistor is proportional to the thickness of the material and inversely proportional to the cross-sectional area across which the heat is being transferred: $R_{vertical} = t/(k{\cdot}A)$, where k is thermal conductivity of the material. Lateral thermal resistor is essentially the constriction or spreading thermal resistance for heat to diffuse laterally from the block into the other parts of the material, and can be calculated by methods in [11]. Thermal capacitor, on the other hand, is proportional to both thickness and area: $C = \alpha{\cdot}c_p{\cdot}\rho{\cdot}t{\cdot}A$, where c_p and ρ are specific heat and density of the material, respectively. Notice that the thermal capacitor derived here is using a single-lumped model instead of a distributed model. Therefore, a scaling factor $\alpha = 0.5$ for thermal capacitances is needed to solve this problem. This scaling factor is derived analytically in [12] for single-lumped vs. distributed electrical RC circuit. It also applies to our thermal RC circuit. It is useful to notice that the derivation methods of thermal Rs and Cs for the primary heat flow path allows us to use the same modeling approach at different level of granularity.

3.3 Secondary Heat Flow Path

The thermal model for the secondary heat flow path is divided into two parts: one corresponding to the interconnect layers, and the other for the path from I/O pads to the printed-circuit board (see Figure 3(a) and (b)).

3.3.1 Interconnect Thermal Model

There are two aspects considered in the interconnect thermal model— 1) Self-heating power of an individual metal wire, which is $P_{self} = I^2{\cdot}R$, where I is the current flowing through the metal wire, $R = \rho_m{\cdot}l/A_m$ is the electrical resistance of the metal wire, ρ_m is the metal resistivity (which is temperature dependent), l and A_m are length and cross-sectional area of the individual wire. Because the interconnect thermal model needs to *predict* wire temperatures before physical layout is available, this means the model has to be able to predict the average wire length of each metal layer. It also needs to be able to predict average current for wires in each metal layer. 2) Equivalent *thermal* resistance for each metal wire and its surrounding inter-layer dielectric. Vias also play an important role in heat transfer among different metal layers, and therefore should be included in the model.

We solve the first aspect of the interconnect thermal model by adopting and extending the statistical *a priori* wire-length distribution model in [13]. This model is developed by Davis *et al.* and is based on Rent's Rule: $T = k_r N^{p_r}$, where k_r and p_r are called Rent's Rule parameters, N is the number of gates in a circuit, T is the predicted number of I/O terminal in the circuit. The Davis model has three wire-length regions—local, semi-global and global. It predicts the number of wires at any specific length. This is called the interconnect density function $i(l)$, where l is the wire length in gate pitches. Figure 5 shows an example wire-length distribution based on ITRS data for high-performance designs at 50nm technology node, where L_{loc}, L_{semi}, L_{glob} are maximum local, semi-global and global wire lengths, respectively. Using interconnect density function $i(l)$, we can calculate the average length and number of wiring nets for each region: (We will use semi-global region as the example throughout Section 3.3.1.)

$$l_{semi} = \chi\,\text{f.o.}\frac{\int_{L_{loc}}^{L_{semi}} i(l)\cdot l\,dl}{\int_{L_{loc}}^{L_{semi}} i(l)\,dl}$$

[1] We have developed a tool to do this job, it will be integrated into the thermal model in the near future.

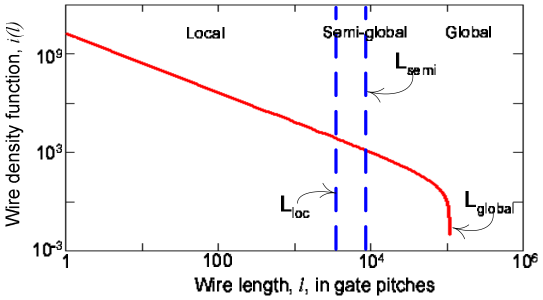

Figure 5: An example of wire-length distribution at 50nm technology node. (adapted from [16])

$$n_{semi} = \frac{1}{\text{f.o.}} \int_{L_{loc}}^{L_{semi}} i(l)\,dl$$

where χ is the correction factor that converts the point-to-point interconnect length to wiring net length (using a linear net model $\chi = 4/(\text{f.o.}+3)$), f.o. is the average number of fan-outs per wiring net. More details can be found in [13].

Next, we calculate the average self-heating power per wiring net: First, total current of the circuit I_{total} is equal to $Power/V_{dd}$. This current flows through all three regions, therefore $I_{glob} = I_{semi} = I_{loc} = I_{total}$. Second, average self-heating power per wiring net can be calculated by

$$P_{self\_semi} = I^2_{single\_net} \cdot R_{single\_net} = \left(\frac{I_{semi}}{n_{semi}}\right)^2 \cdot \rho_m \frac{l_{semi}}{A_{m\_semi}}$$

Last, we calculate self-heating power for each layer of the circuit. "Circuit" here means a circuit block at the interested level of granularity. If, for example, the semi-global region consists of metal layers 4-6, we can say that the number of wires in each layer is $n_{m4} = n_{m5} = n_{m6} = n_{semi}/3$, and

$$P_{self\_m4,m5,m6} = P_{self\_semi} \cdot n_{m4,m5,m6}$$

So far, we are done with the first aspect of interconnect thermal modeling—self-heating power calculation. Next, we calculate the equivalent thermal resistance of wires and its surrounding dielectric.

We first start from a simplistic case. Figure 6(a) shows a single interconnect surrounded by inter-layer dielectric. On top of it and beneath it are interconnects in neighboring layers. d is the thickness of the inter-layer dielectric, W and H are width and height of the interconnect cross section. We try to find thermal resistor R_0, so that $2R_0$ is the equivalent thermal resistance we want. The rectangular cross section of the wire can be approximated by a circle of the same area. Heat is spreading from the wire into the dielectric, the isothermal surface is a cylindrical surface marked by the dashed circle. Equivalent resistance R_0 has to taken into account the top half volume of the shaded cylinder. Using calculus, we get

$$R_0 = \ln\left(\frac{d+2r}{2r}\right)/(\pi \cdot k_{ins} \cdot l)$$

where $r = \sqrt{WH/\pi}$ is the equivalent radius of the wire, l is the length of the wire, and k_{ins} is thermal conductivity of the inter-layer dielectric. (See derivation details in Appendix.)

Figure 6(b) shows the real case: multiple wires are in the same layer. Wire pitch is denoted by D. A phenomenon called thermal coupling happens when neighboring wires dissipate power at the same time. Thermal coupling leads to less effective heat conducting area and change the shape of the isothermal surface. Actual isothermal surface is shown by dashed area in the figure. In this case, each wire's effective heat spreading angle is approximately $\theta = 2 \cdot \arctan(D/(d+H))$, and the corresponding equivalent thermal resistance for each wire is

$$R_0 = \ln\left(\frac{d+2r}{2r}\right)/(\theta \cdot k_{ins} \cdot l)$$

Inter-layer heat transfer can also happen through vias. In our model, we assume that each metal wire has two vias, one connected to upper metal layer, and the other one connected to lower metal layer or device layer in case of metal 1. Thermal resistance of each via can be calculated by $R_{via} = t_v/(k_v A_v)$, where k_v is thermal conductivity of via-filling material. t_v and A_v are thickness and cross-sectional area of the via.

All thermal resistors of wires and vias inside one layer can be considered parallel to each other. Thus, combining thermal resistors of wires and vias in one layer (e.g. metal 4 in semi-global region) of the circuit, we have—

$$R_{m4} = \frac{2R_0}{n_{m4}} \,||\, \frac{R_{via}}{n_{m4}}$$

We are almost done with the interconnect thermal modeling. One last step is to stack the thermal resistors for each layer to construct the whole thermal circuit for all interconnect layers. Currently, the interconnect thermal model doesn't include thermal capacitors, because people usually are more interested in steady-state interconnect temperatures for electromigration and power-grid IR drop analyses. But thermal capacitors can be easily added using the methods presented in Section 3.2 and in this section.

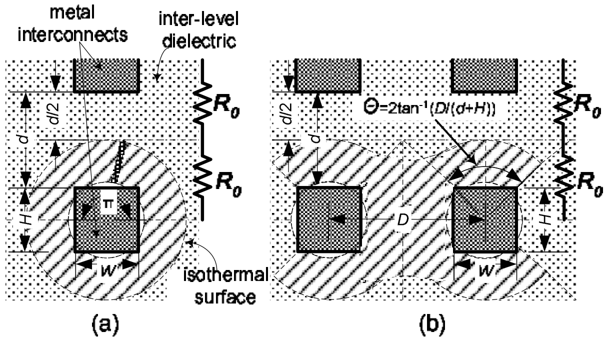

Figure 6: Interconnect structures—(a) stacked single wires (b) real wire structure with multiple wires in each layer.

3.3.2 Thermal Model from I/O Pads to PCB

Our thermal model for the path from I/O pads to PCB consists of a series of thermal RC pairs, each of which represents the thermal resistance and capacitance of pad-bumps/underfill, ceramic substrate, ball/lead array, and PCB convection (see Figure3(b)). As for the derivation, Rs and Cs can be calculated in a similar way as in Section 3.2. The Rs and Cs for the pads/underfill level can be modeled at the interested level of granularity. One ends of these Rs for pads/underfill are connected to the interconnect-level thermal model, the other ends are joined as one node, which is then connected to the RC pair representing ceramic substrate, and so on.

3.4 Speed of the Compact Thermal Model

So far, we have shown all parts of the compact thermal model. The model is derived in a straightforward way and is very computationally efficient. Table 1 shows the computation times of our thermal model to obtain steady-state solutions at different granularities. This computational efficiency means there is virtually no computation overhead for existing CAD tools to integrate the compact thermal model for temperature-aware design.

# of grids	execution time (ms)
5x5	0.12
50x50	2.48
100x100	9.98
160x160	25.8

Table 1: Computation times of our model for steady-state temperatures.

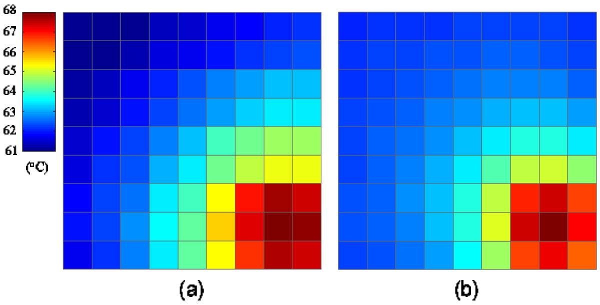

Figure 7: Primary heat flow path: steady-state validation—(a) testing chip measurements (b) our model results with absolute percentage error less than 5%. (In this experiment, lower-right 3x3 dissipaters are turned on with 50W/cm^2 power density.)

4. MODEL VALIDATION

We validate the compact thermal model in the same sequence we derive it—primary heat flow path first, followed by the secondary heat flow path.

4.1 Primary Heat Flow Path

This part of the model is validated against a commercial thermal testing chip[14]. The thermal testing chip has a 9x9 grid of power dissipaters, which can be turned on or off individually. The testing chip is able to measure both steady-state and transient temperatures for each of the dissipaters. It can also be packaged in different kinds of thermal packages. We build the same 9x9 grid-like chip structure in our thermal model. In this experiment, we neglect the secondary heat flow path, because the test chip is plugged in a plastic socket that has very low thermal conductivity. We then turn on sets of power dissipaters in the test chip and assign same power values at the same locations in our thermal model.

	steady-state	transient
average absolute error	1.46%	2.26%
error range	−3.35%—+4.75%	−7.0%—+6.7%

Table 2: Percentage error values for primary heat flow path validations

Figure 7 shows the steady-state thermal plots using measurements from testing chip and results from our thermal model. Transient temperature data from the thermal model are also compared with the testing chip transient measurements, as shown in Figure 8. Table 2 shows the percentage error values, which are calculated by $(T_{model} - T_{chip})/(T_{chip} - T_{ambient})$. Power density in this experiment is 50W/cm^2 in the heat dissipating area. As can be seen, our thermal model of the primary heat flow path is reasonably accurate, with the worst case percentage error values for steady-state temperatures and transient temperatures less than 5% and 7%, respectively.

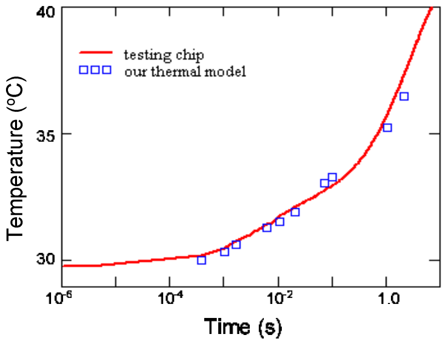

Figure 8: Primary heat flow path: transient validation. Percentage error is less than 7%. (Transient temperature response at one of the dissipaters is shown in this figure.)

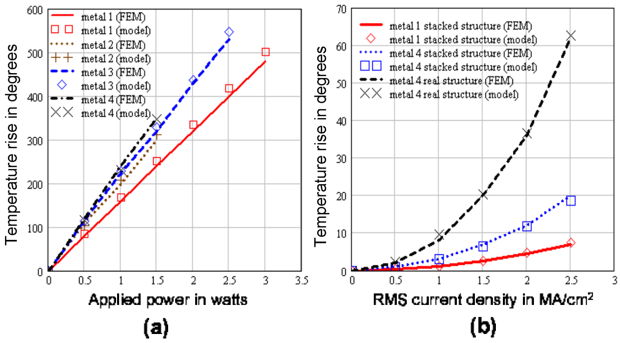

Figure 9: Interconnect thermal model validation—FEM results (lines) from [15], and our thermal model results (markers): (a) stacked single wires—powers are applied to each wire (b) RMS current densities are applied to both test structures.

4.2 Secondary Heat Flow Path

For validation of the interconnect thermal model, we compare our model to the finite-element models (FEM) published in [15]. In [15], the authors build two interconnect test structures in FEM analysis software: one with individual metal wires on top of each other (this corresponds to the case of Figure 6(a)); and the other one with multiple metal wires within each layer(this corresponds to the case of Figure 6(b)). Both test structures have four metal layers at 0.6μm technology. We use exactly the same settings for our interconnect thermal model as in [15] for both test structures, and performe the same two experiments as in [15]—1) For the stacked single-wire test structure, apply different power for each wire and obtain the temperature rise with respect to ambient temperature; 2) For both test structures, we apply different current density for each layer and obtain the temperature rise. Results are shown in Figure 9(a) and (b). As can be seen, Results of our interconnect thermal model match FEM simulation results very well.

For validation of the thermal model from I/O pads to printed-circuit board, there is no straightforward existing data for comparison. But based on the validation of other parts of the thermal model, we have enough confidence that our model for this part is reasonably accurate. A simple calculation using our model based on the thermal specifications of the PowerPC603 CBGA package[10] shows that about 17.5% of total heat is dissipated through the secondary heat flow path.

5. A CASE STUDY

In this section, we present a microprocessor design at a future 50nm technology node as a case study. This case study demonstrates the application of our compact thermal model and the importance of using temperature as a guideline during design. Technology specifications used in this case study are shown in Table 3, the second column of which is taken from [1] and [16]. We also use on-die level-one data cache of Alpha21364 processor scaled to 50nm technology node as an example of localized heating. The scaling process is a linear scaling from known data at 130nm technology, with proper considerations for leakage power and area. Power consumption values of functional units are extracted from a technology-scaled version of Wattch[17].

physical parameters	across die	L1 D-cache
number of transistors	2200 million	70 million
Rent's parameters	$p_r = 0.6$, $k_r = 4.0$	$p_r = 0.6$, $k_r = 4.0$
feature size	50nm	50nm
wiring levels	9	9
area	3.10cm^2	0.08cm^2
power dissipation	218W	52.8W
power density	70.3W/cm^2	660W/cm^2

Table 3: A microprocessor example—across-die vs. L1 D-cache (based on ITRS 50nm technology node[1] and [16]).

We first show that at die-level, using estimated temperature from our thermal model offers much better design estimations for power, delay and interconnect reliability than just using room temperature (20°C) or worst-case temperature, as can be seen from the results presented in Table 4. It is obvious that using room temperature or worst-case temperature yields more errors, therefore leading to possibly wrong design decisions and longer design convergence time.

The second experiment is to show the importance of being able to estimate temperatures at different granularities. This is because different stages of the design process need different granularitiesy of power, delay or reliability estimations, hence different granularities of temperature estimations. By changing the number of grids, i.e. the level of granularity in our thermal model, we can calculate the average temperature across the die, average temperature of the L1 data cache, and max/min temperatures within the L1 D-cache. As can be seen in Table 5, local hot spot like L1 D-cache can have significantly higher temperature than average die temperature. Even within L1 D-cache itself, there is also noticeable temperature gradients. Therefore, during the design of local circuits like L1 D-cache, using average die temperature yields inaccurate design estimates.

	model	room temp.	worst-case temp.(149°C)
leakage power	1.0	0.61	2.85
delay	1.0	0.83	1.25
reliability	1.0	37.40	0.027

Table 4: Temperature estimates using room temperature and worst-case temperature, normalized to the temperature estimates from the thermal model.

	across die	L1 D-cache	within L1 D-cache
temperature (°C)	72.8	117	109-119

Table 5: Temperatures at different levels of granularity.

6. CONCLUSION

In this paper, we have shown that thermal design will be one of the major challenges for the CAD community for sub-100nm designs. To address this challenge, we introduce the idea of *temperature-aware* design, which uses temperature as a guideline during the design flow. We also propose a compact thermal model for temperature-aware design. Results from our thermal model show that temperature-aware design approach can provide more accurate design estimations, and therefore better design decisions and faster design convergence.

7. APPENDIX

We first calculate the thermal resistance of the dark shaded slice of inter-layer dielectric shown in Figure 6(a). It can be written in the form of the integral

$$dR = \int_0^{d/2} \frac{1}{k_{ins}} \frac{dx}{(r+x)d\theta \cdot l} = \frac{1}{k_{ins} \cdot l \cdot d\theta} \ln(\frac{d+2r}{2r})$$

where x is the integral variable, k_{ins} is the thermal conductivity of the inter-layer dielectric, and θ is the angle of the slice.

If we define thermal conductance G as the reciprocal of thermal resistance R, we have

$$dG = k_{ins} \cdot l \cdot d\theta \cdot \frac{1}{\ln(\frac{d+2r}{2r})} \Rightarrow G = \int_0^{\pi} dG = \pi \cdot k_{ins} \cdot l \cdot \frac{1}{\ln(\frac{d+2r}{2r})}$$

so the total equivalent thermal resistance is

$$R = \frac{1}{G} = \ln(\frac{d+2r}{2r})/(\pi \cdot k_{ins} \cdot l).$$

Similarly, thermal resistance for the case of Figure 6(b) can also be derived.

8. REFERENCES

[1] The international technology roadmap for semiconductors (ITRS), 2001.

[2] K. Roy, S. Mukhopadhyay, and H. Mahmoodi-Meimand. Leakage current mechanisms and leakage reduction techniques in deep-submicrometer CMOS circuits. *Proceedings of the IEEE*, 91(2):305–327, February 2003.

[3] T. Kam, S. Rawat, D. Kirkpatrick, R. Roy, G. S. Spirakis, N. Sherwani, and C. Peterson. EDA challenges facing future microprocessor design. *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, 19(12):1498–1506, December 2000.

[4] K. Torki and F. Ciontu. IC thermal map from digital and thermal simulations. In *Proceedings of the 2002 International Workshop on THERMal Investigations of ICs and Systems (THERMINIC)*, pages 303–08, Oct. 2002.

[5] K. Skadron, M. R. Stan, W. Huang, S. Velusamy, K. Sankaranarayanan, and D. Tarjan. Temperature-aware microarchitecture. In *Proc. ISCA-30*, pages 2–13, June 2003.

[6] T-Y. Wang and C. C-P. Chen. 3-D thermal-ADI: A linear-time chip level transient thermal simulator. *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, 21(12):1434–1445, December 2002.

[7] C. J. M. Lasance. Two benchmarks to facilitate the study of compact thermal modeling phenomena. *Components and Packaging Technologies, IEEE Transactions on*, 24(4):559–565, December 2001.

[8] W. Batty et al. Global coupled EM-electrical-thermal simulation and experimental validation for a spatial power combining MMIC array. *IEEE Transactions on Microwave Theory and Techniques*, pages 2820–33, Dec. 2002.

[9] D. Brooks and M. Martonosi. Dynamic thermal management for high-performance microprocessors. In *Proc. HPCA-7*, pages 171–182, January 2001.

[10] J. Parry, H. Rosten, and G. B. Kromann. The development of component-level thermal compact models of a C4/CBGA interconnect technology: The motorola PowerPC 603 and PowerPC 604 RISC microproceesors. *Components, Packaging, and Manufacturing Technology–Part A, IEEE Transactions on*, 21(1):104–112, March 1998.

[11] S. Lee, S. Song, V. Au, and K. Moran. Constricting/spreading resistance model for electronics packaging. In *Proc. AJTEC*, pages 199–206, March 1995.

[12] H. B. Bakoglu. *Circuits, Interconnections, and Packaging for VLSI*. Addison-Wesley Publishing Company, Reading, Massachusetts, 1990.

[13] J. A. Davis, V. K. De, and J. D. Meindl. A stochastic wire-length distribution for gigascale integration (GSI)—part I: Derivation and validation. *Electron Devices, IEEE Transactions on*, 45(3):580–589, March 1998.

[14] V. Székely, C. Márta, M. Renze, G. Végh, Z. Benedek, and S. Török. A thermal benchmark chip: Design and applications. *Components, Packaging, and Manufacturing Technology–Part A, IEEE Transactions on*, 21(3):399–405, September 1998.

[15] S. Rzepka, K. Banerjee, E. Meusel, and C. Hu. Characterization of self-heating in advanced VLSI interconnect lines based on thermal finite element simulation. *Components, Packaging, and Manufacturing Technology–Part A, IEEE Transactions on*, 21(3):406–411, September 1998.

[16] K. Banerjee, S. J. Souri, P. Kapur, and K. C. Saraswat. 3-D ICs: A novel chip design for improving deep-submicrometer interconnect performance and systems-on-chip integration. *Proceedings of the IEEE*, 89(5):602–633, May 2001.

[17] D. Brooks, V. Tiwari, and M. Martonosi. Wattch: A framework for architectural-level power analysis and optimizations. In *Proc. ISCA-27*, pages 83–94, June 2000.