Compact Thermal Modeling for Temperature-Aware Design

Wei Huang[†], Mircea R. Stan[†], Kevin Skadron[‡], Karthik Sankaranarayanan[‡]
Shougata Ghosh[†‡], Sivakumar Velusamy[‡]
Departments of [†] Electrical and Computer Engineering, and [‡] Computer Science, University of Virginia, Charlottesville
{whuang, mircea, sg7w}@virginia.edu, {skadron, karthik, sv7d}@cs.virginia.edu

ABSTRACT

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

Categories and Subject Descriptors:
B.7.2 [Hardware]: Design Aids
J.6 [Computer-Aided Engineering]: Computer-Aided Design.

General Terms: Design, Algorithms.

Keywords: temperature-aware design, temperature-aware computing, thermal model, power-aware design, leakage, reliability.

1. INTRODUCTION

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

Temperature has significant impacts on microelectronic designs—first, transistor speed is slower at higher temperature because of the degradation of carrier mobility. Second, the temperature dependance of leakage power is significant. Leakage power can be orders of magnitude greater at higher temperatures[2]. Third, the interconnect metal resistivity is also dependent on temperature. For example, the resistivity of copper increases by 39% from 20°C to 120°C. Higher resistivity causes longer interconnect RC delay, and hence performance degradation. Last, but not least, reliability is strongly related to temperature. A first order model for the impact of temperature on reliability is the Arrhenius equation: $MTF{=}MTF_0\ exp(E_a/k_bT)$, where T is operating temperature. It is obvious from this equation that increasing the temperature will exponentially decrease the mean time to failure, hence the life time. In summary: for future designs, higher operating temperature will have significant negative impacts on performance, power consumption, and reliability.

Based on the above facts, thermal design is 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 typically use worst-case or room temperature when needed. This can lead to significant estimation errors and hence wrong decisions and longer design convergence time, as can be seen from the 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 upfront 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 estimations of power, performance and reliability. In addition, thermal effects are influenced by placement and routing; for example, putting two hot blocks adjacent to each other will exacerbate the hot spots, while surrounding a hot block by several colder blocks will actually help in cooling down the hot spot. Temperature should thus be included in the cost function in order to achieve optimal placement and routing in sub-100nm. Temperature can also affect manufacturability in terms of packaging and choices of process if the design is thermally limited. Fig. 1 shows a simplified ASIC design flow adapted to become temperature-aware. Temperature profiles are needed at both functional-block level and standard-cell level during the ASIC design flow. Similar arguments also apply to microprocessor and SoC design flows.

From above, we see that it is very important to be able to estimate temperature at different granularities and at different design stages, especially early in the design flow. The estimated temperature can then be used to perform power, performance, and reliability analyses, together with placement, packaging design, etc. As a result, all the decisions use temperature as a guideline and the design is intrinsically 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 is properly considered during the *entire* design flow instead of being determined only after the fact at the end of the design flow. There are a few examples of 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. This work is useful, but the design flow therein cannot be termed as a proper temperature-aware de-

DAC2004 June 7–11, 2004, San Diego, California, USA.

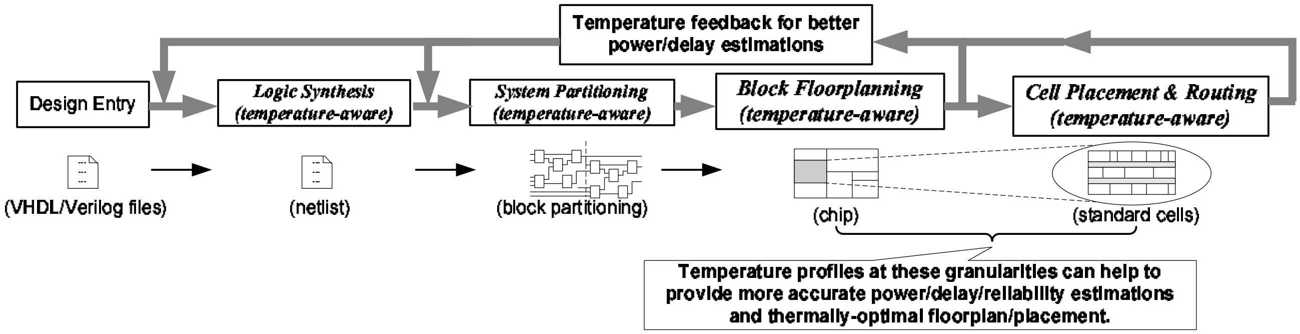
Temperature feedback for better power/delay estimations
Design Entry
Logic Synthesis (temperature-aware)
System Partitioning (temperature-aware)
Block Floorplanning (temperature-aware)
Cell Placement & Routing (temperature-aware)
(VHDL/Verilog files)
(netlist)
(block partitioning)
(chip)
(standard cells)
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.

sign since 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 restart from the beginning if it turns out to be thermally limited.

3. A COMPACT THERMAL MODEL

The first key element for a temperature-ware design methodology is a thermal model to estimate operating temperatures. Fig. 2 shows 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, and, at that point, temperature-aware power estimation is achieved. Similarly, temperature-aware performance and reliability estimations can also be achieved.

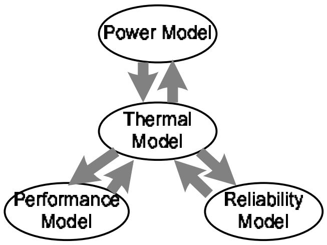

Figure 2: Interactions among thermal model and power, performance and reliability models.

There are a number of existing thermal models for different parts of a microelectronic design. For example, our previous work [5] [6] presents a dynamic compact thermal model, *HotSpot*, only at the microarchitecture level. [7] presents a chip-level thermal model based on full-chip layout. In [8], the authors present package thermal models. In [9], 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 granularities. 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 should be able also to model transient temperature changes. Of course, the model needs to be reasonably accurate to provide useful temperature estimates.

In this paper, we propose a *compact* thermal model that meets all the above requirements and can be used to achieve temperature-aware design. This compact thermal model is an extended version of *HotSpot*, which was proposed in [5] and [6].

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 a similar set of differential equations (there is no thermal equivalent of electrical inductance though). The compact thermal model we propose is essentially a thermal RC circuit. Each node in the circuit corresponds to a block at the desired level of granularity. Heat dissipation of each block is modeled as a current source connected to the corresponding node. Solving this thermal RC circuit gives the temperatures of each node.

Fig. 3(a) shows a modern single-chip CBGA package [10]. Heat generated from the active silicon device layer is conducted through the silicon die to the thermal 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 path and the on-chip interconnect layer. This is because detailed temperature profiles of these parts are 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. Fig. 3(b) shows the thermal RC circuit structure that corresponds to Fig. 3(a). Next, we present the modeling details of each layers along both heat flow paths.

3.2 Primary Heat Flow Path

Fig. 4(a) shows an example thermal circuit of a silicon die with only three microarchitecture blocks from our previous work [5]. We extend the thermal model for the primary heat flow path in [5] by making the model grid-like, thus being able to model temperatures at arbitrary granularities. Fig. 4(b) shows our modeling approach with the granularity of 3x3 grid cells. Each silicon grid cell can be of arbitrary aspect ratio and size, which are determined by the desired level of granularity. We also add to the model a layer of thermal interface material that is absent in [5]. As another small change compared to previous work, the part of the heat spreader that is right under the interface material, as well as the interface material itself, are divided into the same number of grid cells as the silicon die in order to improve accuracy. Other parts in the primary heat flow path are modeled in a similar way as in [5]— the remaining part of the heat spreader is divided into four trapezoidal blocks. The heat sink is divided into five blocks: one corresponding to the area right under the heat spreader and four trapezoids for the periphery. Each grid cell maps to a node in the thermal circuit, and there are vertical and lateral thermal resistors connecting the nodes. Each node also has a thermal capacitor connected to the ambient. The power dissipated in each silicon grid cell is modeled as a “current source” connected to the corresponding node. The package-to-air thermal resistor is calculated from specific heat-sink

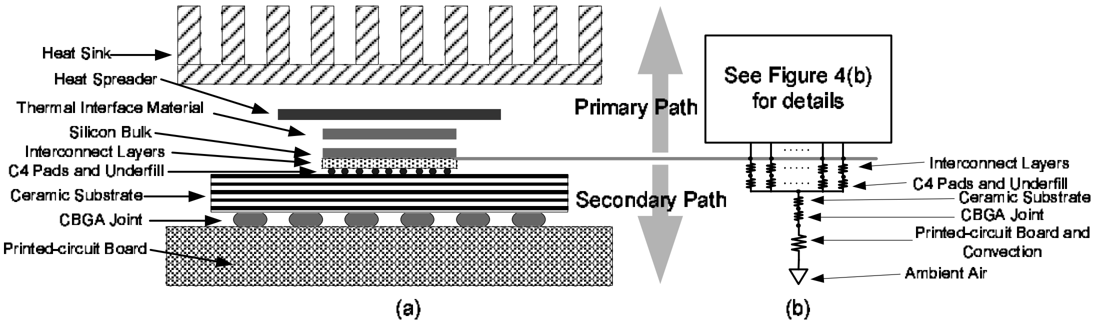

Figure 3: (a) A typical flip-chip, CBGA package with heat sink (adapted from [10]). (b) Corresponding thermal circuit in our thermal model. Thermal capacitors connecting each node to ambient are not shown for clarity.

configurations and ambient conditions. [1]

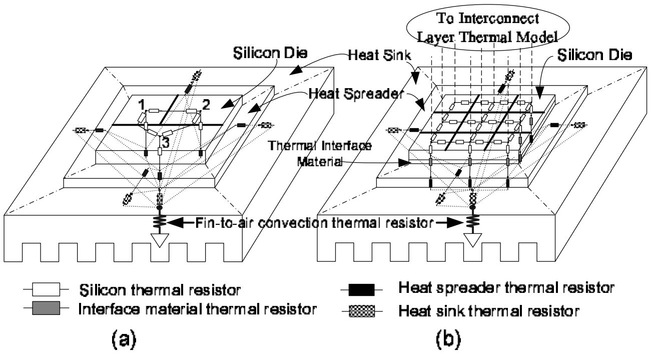
To Interconnect Layer Thermal Model
Silicon Die
Heat Sink
Silicon Die
Heat Spreader
1
2
3
Thermal Interface Material
Fin-to-air convection thermal resistor
Silicon thermal resistor
Interface material thermal resistor
Heat spreader thermal resistor
Heat sink thermal resistor
(a)
(b)

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 grid cells, 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 resistors are proportional to the thickness of the material and inversely proportional to the cross-sectional areas across which the heat is being transferred: $R_{vertical} = t/(k \cdot A)$, where k is thermal conductivity of the material. Lateral thermal resistors are essentially the constriction or spreading thermal resistances for heat to diffuse laterally from one block into other parts of the material, and are calculated by a method described in in [11]. Thermal capacitors, on the other hand, are proportional to both thickness and area: $C = \alpha \cdot c_p \cdot \rho \cdot t \cdot A$, where c_p and ρ are the specific heat and density of the material, respectively. Notice that the thermal capacitor used here is a single-lumped model instead of a more detailed distributed model. Therefore, a scaling factor $\alpha \simeq 0.5$ for thermal capacitances is used to correct this, similar to what was derived analytically in [12] for single-lumped vs. distributed electrical RC circuits. It is useful to note 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 levels of granularity.

3.3 Secondary Heat Flow Path

The secondary heat transfer path helps to remove a non-negligible amount of total generated heat (up to 30%). Neglecting this heat transfer path will lead to inaccurate temperature predictions. In addition, in order to model temperature-affected on-chip interconnect delay and life time, the thermal model of the interconnect metal layers is needed, which is part of the secondary heat transfer path. In this paper, 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 the I/O pads to the printed-circuit board (see Fig. 3(a) and (b)).

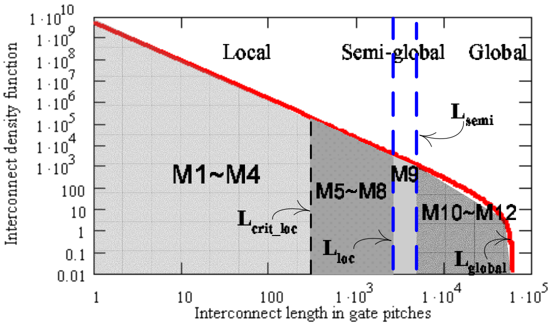

Figure 5: An example of wire-length distribution at 45nm technology node with 12 layers of metal. Shaded areas correspond to assigned metal layers for wires with different length.

3.3.1 Interconnect Thermal Model

There are two aspects considered in the interconnect thermal model: 1) the self-heating power of an individual metal wire, which is $P_{self} = I^2 \cdot R$, where I is the current flowing through the wire, $R = \rho_m \cdot l / A_m$ is the electrical resistance, ρ_m is the metal resistivity (which is temperature dependent), l and A_m are the 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 on each metal layer. It also needs to be able to predict the 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 also be included in the model. In this paper, we only briefly introduce the interconnect thermal model due to the limited space. More details can be found in the extended technical report based on this paper [13].

We solve the first aspect of the interconnect thermal model (self heating) by adopting and extending the statistical *a priori* wire-length distribution model in [14]. This model is is based on Rent's Rule: $T = k_r N^{p_r}$, where k_r and p_r are 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. It is important to note that an interconnect thermal model at high levels of abstraction strongly depends on the *a priori* wire-length distribution model, and hence is limited by the accuracy and efficiency of the wire-length distribution model. Three wire-length regions are considered in [14]—local, semi-global and global. The model predicts the number of wires of any specific length, which is called the interconnect density function $i(l)$, where l is the wire length in gate pitches. Fig. 5 shows an example wire-length distribution based on ITRS data [1] for high-performance designs at the 45nm technology node with 12 layers of metal interconnect, where L_{loc}, L_{semi}, L_{glob} are maximum local, semi-global and global wire lengths, respectively. Using the

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

interconnect density function $i(l)$, we calculate the average length and number of wiring nets for each region. For example, using the semi-global region:

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

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

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 [14].

Once the wire length distribution in each region is known, we assign the interconnects with different lengths to different metal layers. For interconnects that are predicted to be long by the model, repeaters are needed in order to achieve minimum delay. The critical wire-length between repeaters (L_{crit}), the delay for one section of buffered interconnect (τ_{crit}), the optimal number of repeaters (Nr_{crit}) and optimal size of repeaters (s_{crit}) for interconnects in each region can be found using the repeater insertion model proposed in [15]. With the information about repeater insertion, all the interconnects can then be assigned to different metal layers. An example of metal layers assignment corresponding to the design in Fig. 5 is also shown in the same figure — local interconnects that are shorter than L_{crit_loc} are assigned to metal layers 1 through 4; local interconnects that need repeaters are assigned to metal layers 5-8; semi-global interconnects are assigned to metal 9 and global interconnects are assigned to metal layers 10 through 12.

In order to calculate the average self-heating power per interconnect in each metal layer, the average current that causes wire self-heating needs to be calculated first. For each switching event, half of the energy drawn from the power supply is dissipated in the form of heat on the charging/discharging transistor and on the output interconnect, we have

$$I_{avg}^2 (R_{tr} + R_{wire}) t_d = \frac{1}{2} \alpha\, C_L V_{dd}^2$$

From this equation we find the average self-heating current I_{avg} per wire in each metal layer. R_{tr} is the on-resistance of the transistor, R_{wire} is the wire resistance, α is the switching activity factor, C_L is the load capacitance. Average values for these are calculated from ITRS data [1] and the repeater insertion model [15]. The delay of the switching event, t_d, is approximated as as τ_{crit} for interconnects with repeaters, and $clock\_cycle\_time/logic\_depth$ for interconnects without repeaters.

The above wire length and average current calculations based on [14] and [15] are only valid for signal interconnects. Wire length and currents for the power supply grid, namely V_{dd} and GND, have not been considered yet. We do that by building a grid-like resistive network model for the power and ground, somewhat resembling the thermal circuit used for modeling the primary heat flow path in Section 3.2. Each resistor connecting two nodes in the same metal layer is now the electrical resistance of one power supply grid section. Resistors connecting power grid nodes of different metal layers represent the vias. The topology of the network is obtained by knowing the pitch between power rails in each metal layer, average length and number of power grid sections between power grid. Next, by applying currents to the top-layer nodes that are at the C4 pads sites, the resistive network is solved to find the average self-heating current of the power grid in each metal layer.

With all the above information of average interconnect length and average current in each layer (for both signal interconnects and power grid sections), we calculate the average self-heating power per interconnect in each metal layer:

$$P_{self} = I_{avg}^2 \cdot R_{wire} = I_{avg}^2 \cdot \rho_m \frac{l_{wire}}{A_{wire}}$$

where A_{wire} and l_{wire} are the cross-sectional area and the average length of interconnects or power grid sections in each metal layer, respectively.

Last, we calculate the self-heating power for each metal layer of the circuit. “Circuit” here means a circuit block at the desired level of granularity. If, for example, the global region consists of metal layers 10 through 12, we calculate the self-heating power of metal 10 as:

$$P_{self\_m10} = P_{self\_glob\_sig} \cdot \frac{n_{glob\_sig}}{3} + P_{self\_glob\_pwr\_net} \cdot n_{pwr\_net\_m10}$$

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 the surrounding dielectric.

We first start from a simplistic case. Fig. 6(a) shows a single interconnect surrounded by inter-layer dielectric. On top and below 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 the thermal resistor associated with each wire R_0; $2R_0$ represents the series connection for the two wires that will be used in the thermal circuit. 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. The equivalent resistance R_0 has to take into account the top half volume of the shaded cylinder. Using calculus, we get:

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

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. (More details can be found in [13].)

Fig. 6(b) shows the real case: multiple wires are in the same layer. The 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. The actual isothermal surface is shown by the 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 becomes:

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

Inter-layer heat transfer can also happen through vias. In our simple model, we assume that each metal wire has two vias, one connected to the upper metal layer, and another one connected to the lower metal layer. This is a simplistic assumption that will need to be refined in the future, but which does not seem to impact the results significantly. The 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 the local region) of the circuit, we obtain:

$$R_{m4} = \frac{2R_{0\_sig}}{n_{m4\_sig}} \,||\, \frac{2R_{0\_pwr\_net}}{n_{m4\_pwr\_net}} \,||\, \frac{R_{via}}{n_{m4\_sig} + n_{m4\_pwr\_net}}$$

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, but these will be added using the methods presented in Section 3.2 and in this section. Designers are usually more interested in steady-state interconnect temperatures for electromigration and power-grid IR drop analyses.

3.3.2 *Thermal Model from I/O Pads to PCB*

Our model for the heat flow 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 Fig. 3(b)). Rs

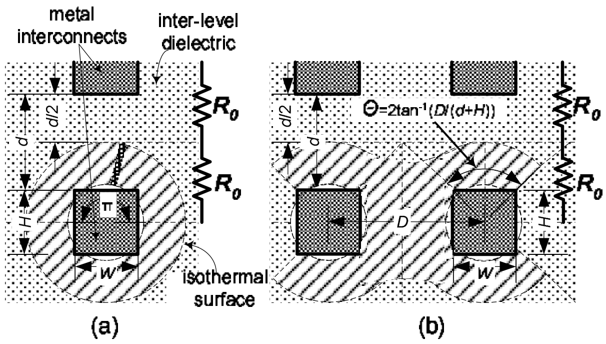

Figure 6: Interconnect structures—(a) stacked single wires (b) real wire structure with multiple wires in each layer.

and Cs are calculated in a similar way as in Section 3.2. The Rs and Cs for the pads/underfill level are modeled at the desired level of granularity. One end for each of these Rs for pads/underfill is connected to the interconnect-level thermal model, the other end is joined into one node, which is then connected to the RC pair representing ceramic substrate, and so on.

3.4 Simulation Speed for the Compact Thermal Model

So far, we have shown all the parts of the compact thermal model. The model is derived in a straightforward way and is 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 design methodologies to integrate the compact thermal model for temperature-aware design. Details of the circuit solver for our model can be found in [13].

# of grid cells	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.

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 test chip [16]. The thermal test chip has a 9x9 grid of power dissipators, which can be turned on or off individually, with an embedded thermal sensor for each grid cell. The test chip can measure both steady-state and transient temperatures for each of the grid cells. We built the same 9x9 grid-like chip structure in our thermal model. In this experiment, we neglected the secondary heat flow path, because the test chip is wire bonded and plugged in a plastic socket that has very low thermal conductivity. We then turned on sets of power dissipators in the test chip and assigned the 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

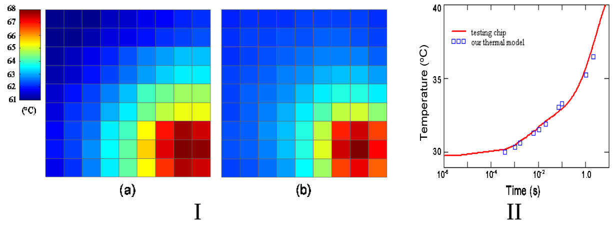

Figure 7: (I)— Steady-state validation of the compact thermal model: (a) Test chip measurements (b) Results from the model with errors less than 5%. (II)— Transient validation of the compact thermal model. Percentage error is less than 7%. (Transient temperature response of one power dissipator is shown here.)

Fig. 7(I) shows the steady-state thermal plots using measurements from the test chip and results from our thermal model. Transient temperature data from the thermal model are also compared with the test chip transient measurements, as shown in Fig. 7(II). Table 2 shows the percentage error values, which are calculated by $(T_{model} - T_{chip})/(T_{chip} - T_{ambient})$. The power density in this experiment is 50W/cm^2 in the heat dissipating area (the 3x3 lower-right corner). As can be seen, our thermal model of the primary heat flow path is reasonably accurate, with the worst case error values for steady-state temperatures and transient temperatures less than 5% and 7%, respectively.

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 [17]. There 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 Fig. 6(a)); and the other one with multiple metal wires within each layer(this corresponds to the case of Fig. 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 [17], and perform the same two experiments—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, apply different current density for each layer and obtain the temperature rise. Results are shown in Fig. 8(a) and (b). As can be seen, the 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]

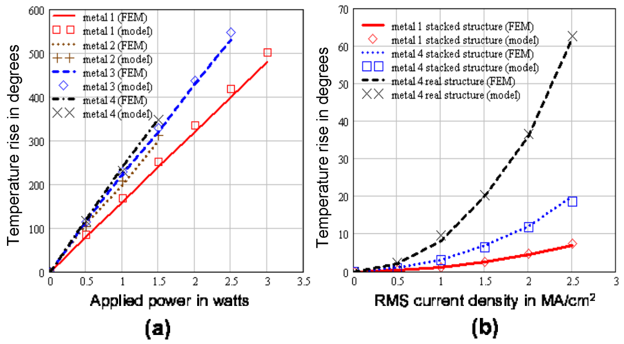

Figure 8: Interconnect thermal model validation—FEM results (lines) from[17], and our thermal model results (markers): (a) stacked single wires—powers are applied to each wire (b) RMS current densities are applied to both test structures.

shows that about 17.5% of total heat is dissipated through the secondary heat flow path.

5. A CASE STUDY

In this section, we briefly present a microprocessor design at a future 45nm technology node as a case study. More details can be found in [13]. 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 [18]. We use an on-die level-one (L1) data cache approximating that of the Alpha 21364 processor scaled to 45nm 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 [19].

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	45nm	45nm
wiring levels	12	12
area	3.10cm^2	9.56mm^2
power dissipation	218W	60.9W
power density	70.3W/cm^2	637W/cm^2

Table 3: A microprocessor example—across-die vs. L1 D-cache (based on ITRS 45nm technology node[1] and [18]).

We first show that at the 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 or worst-case temperature as can be seen from the results presented in Table 4. Simply using room temperature or worst-case temperature yields more errors, therefore leading to possibly incorrect 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 granularities of power, delay or reliability estimations, hence different granularities of temperature estimations. By changing the number of grid cells, 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, a local hot spot like an L1 D-cache can have a significantly higher temperature than the average die temperature. Even within the L1 D-cache itself, there are also noticeable temperature gradients. Therefore, during the design of specific blocks like the L1 D-cache, using average die temperature yields inaccurate design estimates. From the last column of Table 5, we can also see the influence of number of grid cells on the accuracy of maximum L1 D-cache temperature predictions.

As another example of how our compact thermal model can be applied, recent work on a leakage power simulator [20] uses our compact thermal model to predict operating temperature of the microprocessor and hence closes the loop of temperature and leakage power estimation.

6. CONCLUSION

We believe 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 a temperature-aware methodology can provide more accurate design estimations, and therefore better design decisions and faster design convergence.

	model	room temp.	worst-case temp.
leakage power	1.0	0.61	2.85
delay	1.0	0.83	1.25
life time	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.

# of grids (die)	die avg. T	D-cache avg. T	D-cache max T
25x25	72.8	115.4	120.5
30x30	72.8	115.4	123.7
35x35	72.8	115.4	126.7

Table 5: Temperatures at different levels of granularity (°C).

Acknowledgments

This work is supported in part by the National Science Foundation under grant no. CCR-0133634, two grants from Intel MRL, and an Excellence Award from the Univ. of Virginia Fund for Excellence in Science and Technology. We would also like to thank the anonymous reviewers for their insightful comments.

7. REFERENCES

[1] The international technology roadmap for semiconductors (ITRS), 2003.

[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 *Proc. 8th 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] K. Skadron, K. Sankaranarayanan, S. Velusamy, and D. Tarjan, and M. R. Stan, W. Huang, Temperature-Aware Microarchitecture: Modeling and Implementation. *ACM Transactions on Architecture and Code Optimization*, To appear, 2004.

[7] 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.

[8] 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.

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

[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] W. Huang, M. R. Stan, K. Skadron, K. Sankaranarayanan, S. Ghosh, and S. Velusamy. Compact thermal modeling for temperature-aware design. Tech Report CS-2004-13, Univ. of Virginia Dept. of Computer Science, April. 2004.

[14] 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.

[15] R. H. J. M. Otten, R. K.Brayton. Planning for performance. In *Proc DAC 1998*, pages 122–127, 1998.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] Y.-F. Tsai. An Architecture-Level Leakage Power Simulator. Ph.D. Forum at DATE 2004, Feb. 2004.