

An Improved Block-Based Thermal Model in HotSpot 4.0 with Granularity Considerations

Wei Huang[†], Karthik Sankaranarayanan[†], Robert J. Ribando[‡], Mircea R. Stan^{*}, and Kevin Skadron[†]

{whuang@, ks4kk@cs., rjr@, mircea@, skadron@cs.}virginia.edu

Depts. of Computer Science[†], Electrical and Computer Engineering^{*}, and Mechanical and Aerospace Engineering[‡]

University of Virginia

Charlottesville, VA 22904, USA

Abstract

This paper describes our most recent improvements to the HotSpot thermal model. First, in response to a previous paper in WDDD 2006 [1] that cites some accuracy shortcomings, we improve treatment in HotSpot’s “block model” of high aspect-ratio blocks, high power densities, and convection boundary conditions. These improvements are included in the newest release of the HotSpot software, version 4.0.¹ Second, to help guide users’ modeling efforts and choice of modeling resolution, we propose an analytical approach to find the relationship between the power modeling granularity and the associated peak temperature rise at that granularity, i.e. the relationship between the size and peak temperature of a uniform heat source. We theoretically confirm the existence of a strong spatial temperature low-pass filtering effect: a tiny heat source is much cooler than a large heat source with the same power density.

1 Introduction

Along with the continued scaling of CMOS technology, on-chip operating temperature has become a major design challenge. The microarchitecture is especially important, because the architecture definition fixes what subsequent design stages—circuit implementation, packaging, etc.—must accommodate. Temperature-aware design in early, pre-RTL design stages in turn requires a fast yet accurate architectural thermal model to explore large regions of the design space. HotSpot [2] is one such model widely used by the computer architecture research community. Compared to traditional thermal modeling techniques such as detailed finite-element models, HotSpot sacrifices detail and accuracy to achieve compactness and ease of use. To date, HotSpot seems to have been most often used with existing architectural simulation infrastructures such as SimpleScalar² and Wattch [3], but it is designed as a portable library that can be used with a wide range of modeling infrastructures.

HotSpot is available online³. It consists of three major parts: (1) the by-construction parameterized structure of the thermal model itself; (2) a linear solver; (3) the input floorplan, power map and power trace. The outputs of HotSpot are static and transient temperatures of each floorplan unit and package components. HotSpot is efficient due to the fact that it is a compact network of thermal resistors and capacitors with a man-

ageable number of nodes. The node temperatures are readily solved by efficient linear solvers. On the other hand, although the absolute accuracy is inevitably traded off for computational efficiency, HotSpot is still reasonably accurate with respect to detailed finite-element thermal models and real chip thermal measurements, as reported in our previous validations [4, 5, 6].

HotSpot was first introduced in 2003 with only the block model. Later on, thermal interface material (TIM) was added to HotSpot 2.0 to closely model real-world thermal packages. Starting from HotSpot 3.0, we also introduced a regular-grid-based model along with the block-based model. One major reason to develop the grid model is to achieve more accuracy by modeling lateral heat transfer paths in more detail than the block model. Therefore, some of the accuracy issues in the block model that we try to improve in this paper have been automatically taken care of by the grid model, for example, the block aspect ratio issue that we will see shortly.

In a previous WDDD paper, Fetis and Michaud [1] evaluated the accuracy the HotSpot *block-based* model with an open-source detailed finite-element numerical software, FreeFEM3d⁴ and an analytical thermal solver, ATMI⁵. The HotSpot block model was found to be reasonably accurate in most cases. On the other hand, noticeable and even significant errors were found under certain evaluation scenarios. All of these scenarios contain some extreme configurations that are either not fully taken into account by earlier releases of the HotSpot block model (e.g. functional blocks with very high aspect ratios), or uncommon in typical designs (e.g. extremely high power densities). In addition, the assumed heatsink convection boundary condition in [1] is not exactly the same as that in HotSpot 3.0. All of these lead to the reported inaccuracy of HotSpot.

In this paper, we acknowledge and explain these inaccuracies and provide solutions to improve the accuracy of the HotSpot block-based model. To evaluate the accuracy benefits of these improvements, we compare against a popular and reliable commercial detailed finite-element model, ANSYS⁶. By doing this, we also make the HotSpot block model more robust so that it still maintains good accuracy even under the aforementioned extreme conditions. All these improvements are achieved by making as little modifications as possible to the existing HotSpot block model structure so that the efficiency of the HotSpot block-based model is preserved. As a result, a new HotSpot release, HotSpot 4.0, will be available online. This pa-

¹Version 4.0 will be released before ISCA this coming June.

²<http://simplescalar.com>

³<http://lava.cs.virginia.edu/HotSpot/>

⁴<http://www.freefem.org/ff3d/>

⁵<http://www.irisa.fr/caps/projects/ATMI/>

⁶<http://www.ansys.com/>

per is mainly focused on the HotSpot block model, but we also briefly discuss the extent to which our findings are applicable to the HotSpot grid model and how to improve the accuracy of the grid model under various conditions.

In addition to the accuracy improvement of HotSpot, another important issue in architecture-level thermal modeling that has not drawn much attention so far is to find the proper granularity at which temperature is modeled—does temperature need to be modeled for each processor core, each within-core functional block, each individual register file entry, or even each individual gate or transistor? Of course, the answer depends on various factors such as the design level one works at, the available granularity of power estimation, the desired design complexity and accuracy tradeoff, etc. But first of all, in order to get a good handle on this, it is helpful to have some theoretical analysis on how the size of a heat source would affect the peak temperature and the spatial temperature distribution on silicon, based on the physical heat transfer equations. In this paper, we propose such a theoretical analysis, which can greatly help researchers to find the right thermal and power modeling granularity in their thermal-related research. It can also help design engineers to find the right granularity at which thermal monitoring and thermal management techniques are deployed, without missing hot spots on one hand or over-engineering to an unnecessarily fine granularity on the other.

This paper is organized as follows. In Section 2, we first identify the causes of the errors of the HotSpot block-based model reported in [1] and provide corresponding solutions. We then show the resultant improved accuracy of the HotSpot block model by performing experiments similar to those in [1] in Section 3. In Section 4, we present the theoretical analysis of the relationship between the heat source size and the peak temperature, and provide guidelines for choosing the proper thermal modeling and management granularity. The paper is finally concluded by Section 5.

2 Accuracy Improvement of the HotSpot Block Model

2.1 Sources of Inaccuracy

2.1.1 Sources of inaccuracy as mentioned in [1]

In [1], the authors performed evaluate the HotSpot 3.1 block-based model under different scenarios using FreeFEM3d and ATMI. Both FreeFEM3d and ATMI are detailed models with more accuracy and less computational efficiency than HotSpot. Here we quote their main conclusions and briefly explain the source of the observed problems and our solutions for each. Subsequent sections then go into further detail.

1. “HotSpot is sensitive to space discretization. Different floorplans modeling the same power density map may give, with HotSpot, different temperature numbers.”—This is true when the block-based HotSpot model is used. Because each functional block is approximated by only one node in the model, the associated lumped thermal resistors and capacitors cannot fully model the distributed nature of heat transfer. In particular, the error is more noticeable for blocks with high aspect ratios where the lateral heat transfer in one direction dominates the other direction. This

simply requires higher resolution, and the solution is to further divide these high-aspect-ratio blocks into sub-blocks with aspect ratios close to unity. Note that this issue is not a concern for the HotSpot grid model, because blocks are automatically divided into grid cells with aspect ratio close to unity, although the grid model is slower than the block model.

2. “Floorplans with a larger number of blocks seem to be more accurate.”—This is essentially equivalent to having more nodes in the block model to solve and hence a higher-resolution thermal model closer to detailed finite-element models. On the other hand, even in this case, more error can be seen if there are some blocks with extreme aspect ratios. Those blocks also need to be divided into sub-blocks close to squares. Again, this is not a concern for the grid model.
3. “Naive floorplans may give an error exceeding 200%.”—This happens in an extreme case where a floorplan has functional blocks with extreme aspect ratios, e.g. a 10:1 aspect ratio in the experiment in [1] that results in the 200% inaccuracy. As we will see in Section 3.1.2, dividing this block into ten sub-blocks with aspect ratio of 1:1 greatly improves the accuracy.
4. “HotSpot underestimates the slope of the temperature response for small times.”—To investigate this, we performed transient thermal experiments in ANSYS, a popular commercial finite-element analysis software, with the same geometry and boundary conditions as those in [1]. We found discrepancies in HotSpot, but not as severe as FreeFEM3D suggests, which disagrees with ANSYS.⁷ We improve the transient accuracy of HotSpot with respect to ANSYS by using a constant 0.5 scaling factor for lumped thermal capacitors.
5. “HotSpot underestimates the amplitude of time-varying temperature oscillations.”—This is also due to the transient error of HotSpot. By using a constant scaling factor for thermal capacitors, the transient accuracy is improved compared to the reference ANSYS model.

2.1.2 Other sources of inaccuracy within HotSpot

In addition to the above explanations to the remarks in [1], we also notice the following factors affect HotSpot’s accuracy. The first one is dealt with in this paper. The other one is a simple extension for future upcoming HotSpot releases.

1. Different boundary condition assumptions lead to different temperature estimations. For example, at the heatsink/ambient interface, an isotherm condition is assumed in HotSpot, whereas a convection boundary condition is assumed by [1], which is more realistic and leads to non-isotherm temperature distribution at the heatsink/ambient interface. In this paper, we adopt this more realistic convective boundary condition and implement it in HotSpot 4.0 to further improve accuracy.

⁷Although we did not investigate extensively, the source of the FreeFEM3D discrepancy appears to stem from FreeFEM3D’s choice of mesh.

- Another source of inaccuracy in HotSpot comes from the fact that certain material properties, such as thermal conductivity and specific heat, are temperature dependent. Approximating them with constant values thus introduces errors. Although it is fairly straightforward to include this in HotSpot in the form of lookup tables, this is not the focus of this paper and is a topic for future work.

2.1.3 Notes on HotSpot’s accuracy evaluation

When evaluating the accuracy of HotSpot, and in general, all thermal models, it is also important to keep the following factors in mind in order to obtain a fair evaluation.

- Different reference finite-element models do have noticeable differences in temperature results. For example, ANSYS and FreeFEM3d have discrepancies in both steady-state and transient temperature estimations, as will be shown in Section 3. This can be caused by different levels of details of the finite-element models, different types of finite element used, different mesh alignment, different embedded solvers, etc. Therefore, it is important to first make sure that the detailed finite-element model is set up correctly and is used at the proper level of detail. We use ANSYS as our reference.
- The absolute error of HotSpot can be amplified when functional blocks are applied with unrealistic powers or power densities. For example, if a $1\text{mm} \times 1\text{mm}$ square source is applied with 10W of power, which results in a power density of $10\text{W}/\text{mm}^2$ —an extreme value much higher than the peak block power density in state-of-the-art high-performance microprocessors. HotSpot was not designed with such a high power density over a large area in mind. (Some tiny on-chip circuit structures can actually have extremely high power densities. However, due to their small sizes, the peak temperature is not significant. Section 4 discusses this phenomenon in more detail.)
- The granularity of the power model limits the accuracy that the thermal model can best achieve. HotSpot assumes that for each block or grid cell, the power density is applied uniformly over that area. For example, if the power estimations used as inputs to HotSpot are obtained at the granularity of individual architecture blocks. HotSpot cannot give accurate actual temperature distribution within the blocks unless the circuitry therein is very uniform. For example, if only a single power density is applied for the L1 cache, the temperature distribution within the L1 cache will not reflect the layout of banks, decoders, etc., no matter how accurate HotSpot itself is. Using finer resolution will give a detailed temperature gradient across the cache, but the magnitude of that gradient within the cache is a fiction because it does not account for the structural variation within that space. Thus, in most cases, it is unnecessary to perform thermal analysis at any finer granularity than the granularity of the power model.
- Absolute transient accuracy is harder to achieve than static accuracy in HotSpot without introducing significant extra model complexity. This is due to the lumped structure

of HotSpot and the distributed nature of actual transient thermal response. Scaling factors to thermal capacitances can match the thermal time constants between lumped and distributed systems, but cannot guarantee perfect match over the entire transient temperature response. The only way to achieve ultimate transient accuracy is to use a very fine 3-D mesh to model the system, which inevitably introduces significant computational overhead, and is probably not suitable for architecture-level simulations. As will be shown in Section 3.2, using a constant 0.5 capacitance scaling factor in the block model achieves fairly good accuracy with respect to ANSYS for most of the time scales.

2.2 Improvements of the HotSpot Block Model

Having reviewed the above possible sources of inaccuracies, in this section, we list the detailed improvements we have made to the HotSpot block-based model. We will further compare the results from the improved HotSpot model (HotSpot 4.0) with results from HotSpot 3.1, ANSYS and FreeFEM3d in Section 3.

- Block lateral aspect ratio**—It is important to keep the aspect ratio close to unity. Fig. 1(a) shows a functional block with high lateral aspect ratio. Only one node is used to represent its temperature in HotSpot3.1 and earlier releases. The four lumped lateral thermal resistors connected to that node are also shown. In Fig. 1(b), this block is divided into several sub-blocks with close-to-unity aspect ratio. With this modification, the lateral heat transfer within the block is modeled with greater fidelity. As mentioned before, for the grid model, lateral aspect ratio is not a problem.

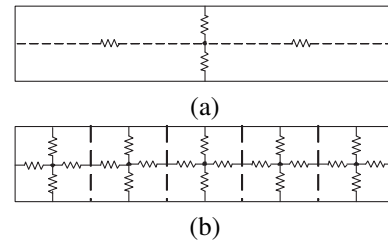


Figure 1. A block with high aspect ratio—(a) in HotSpot3.1, only one node represents the block for computational efficiency. (b) in HotSpot 4.0 model, the block is divided into sub-blocks with aspect ratio close to unity. The lateral heat transfer paths are modeled with more details but more computational complexity.

- Convective boundary condition**—A more realistic convective boundary condition at the heatsink-ambient interface makes HotSpot more accurate. To achieve this, we eliminate the isothermal nodes in the HotSpot 4.0 block model. This is also applicable to the grid model. Fig 2(a) shows the model structure in earlier HotSpot releases, in which the center part of the upper surface of heat spreader is approximated to be isothermal and has only one node (each black dot is a node). The heatsink-ambient interface also has only one node. In the real case, these surfaces are not fully isothermal. Accuracy can therefore be improved by removing the isothermal nodes and modeling the heatsink at the same level of details as the silicon die. Furthermore,

the convection interface between heatsink and ambient air can be modeled with multiple convection surfaces (hence, multiple nodes) with a constant heat transfer coefficient.

$$R_{\text{convec}i} = \frac{1}{hA_i} \quad (1)$$

where $R_{\text{convec}i}$ is the convection thermal resistance for the i th sub-area of the heatsink convection surface, h is the constant heat transfer coefficient, and A_i is the sub-area. The resulting thermal model structure is shown in Fig. 2(b). In earlier HotSpot releases, there is a single convection resistor for the entire heatsink convection area (e.g. $R_{\text{convec}} = 0.1K/W$). The heat transfer coefficient h in (1) can thus be found by solving h from $R_{\text{convec}} = 1/(hA_{\text{tot}})$ to make sure the equivalent total convection thermal resistance in HotSpot 4.0 is the same as before. By doing this, the convection boundary condition in HotSpot is exactly the same as the one used in [1] and is much closer to the real situation. Modeling heatsink with more details introduces more computing overhead to HotSpot. However, as long as the floorplan does not contain too many blocks, the overhead remains tolerable.

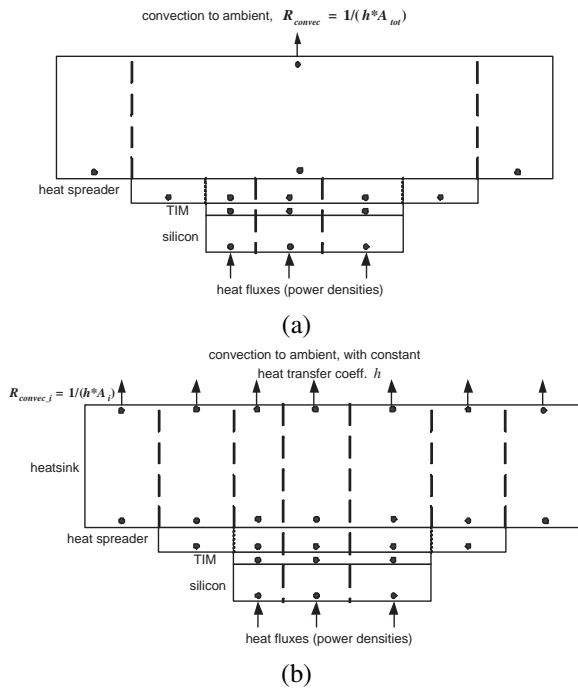


Figure 2. (a) HotSpot 3.1 model structure. There is only one convection resistor from heat sink to ambient air, with top surface of heat spreader and heat sink both assumed to be isothermal. (b) Improved model structure in HotSpot 4.0. The center part of heat sink is modeled at the same level of detail as the silicon. The isotherm nodes are replaced with multiple nodes connected by different convection resistors.

3. *Lateral thermal resistance*—Use a more accurate equation to derive the lateral thermal resistors. In HotSpot, we change to an expression similar to that of the derivation

of vertical thermal resistors

$$R_{\text{lateral}} = \frac{1}{k} \frac{W}{Lt} \quad \text{or} \quad \frac{1}{k} \frac{L}{Wt} \quad (2)$$

where k is the thermal conductivity, W and L are the lateral dimensions of the block, and t is the thickness of the block. When block aspect ratios are closer to unity, this expression turns out to be more accurate than the original expression of spreading/constriction resistance [7] used by earlier HotSpot releases. This is because the discrepancies in the boundary conditions in HotSpot and those in [7] are more serious with small and square blocks. Actually, in HotSpot 3.1, this option to use (2) is available in the function `getr()`. This is also applicable to the grid model.

4. *Transient model*—Use a better scaling factor for thermal capacitors. For transient thermal modeling, since we use lumped thermal capacitors to approximate heat accumulation which has a distributed nature, scaling factors are needed to match the lumped thermal RC time constant to the distributed thermal time constant. Through detailed comparisons of ANSYS and HotSpot transient thermal simulations for the same packaging structure, we found that using a uniform scaling factor of 0.5 for all material layers matches the best. Since the heat spreader and heat sink are both now divided into detailed blocks according to the silicon die, the heat accumulation is already modeled in more detail, so there is no need to include an additional scaling factor to further estimate the lumped effect of the actual thermal mass. Therefore, we remove the additional scaling factors [8] for the silicon and the thermal interface material (TIM) layers in HotSpot 3.1, and replace them all with a single 0.5 scaling factor.

2.3 Some Words about the HotSpot Grid Model

The default method to achieve higher accuracy in HotSpot is to use the grid model provided since HotSpot 3.0, where the silicon die and different package components can automatically be divided into regular grid cells. This is essentially a finite-difference model. Some of the aforementioned issues in the block model such as block lateral aspect ratio and number of nodes are intrinsically dealt with by the grid model. However, the grid model trades off model compactness for accuracy, incurring more computational overhead. Efficient solvers are crucial to improve the speed of grid model. As part of our future work, we will try to make it possible to incorporate existing external fast linear solvers into HotSpot by providing standard interfaces. Other modifications to the block model, such as removing the isothermal nodes and changing the thermal capacitance scaling factor, also help to improve the accuracy of grid model.

It is important to notice that, similar to the lateral aspect ratio issue in the block model, it is vertical aspect ratio can become a problem in the grid model when a very fine grid mesh is used. In those cases, the lateral size of a grid cell can be much less than the layer thickness, resulting in a high vertical aspect ratio, degrading the accuracy of the grid model. In such situations, each layer needs to be further divided into multiple sub-layers to bring the vertical aspect ratio close to unity as well. Additionally, the scaling factor for thermal capacitors also needs

to change since we are modeling each layer with multi-lumped thermal RC ladders instead of one-lumped RC ladder.⁸

3 HotSpot Block Model Improvement Results

With the above modifications to the HotSpot block model, we performed several experiments similar to those in [1] to compare the accuracy of HotSpot 4.0 and HotSpot 3.1 to ANSYS and FreeFEM3d for both steady-state and transient temperature estimations.

As mentioned above, for most of the experiments in this paper, we use ANSYS as our primary reference finite-element model. ANSYS allows users to have a better control on the level of spatial discretization (mesh granularity) and the shape of the finite element (e.g. tetrahedral vs. quadrilateral elements) so that greater accuracy can be achieved with smaller elements. In our ANSYS experiments, we use multiple meshing levels and types of elements, and ensure that the results are consistent across them.

3.1 Steady-State Temperature

3.1.1 EV6 floorplan

The package geometry used in this paper for HotSpot evaluations is similar to Fig. 2. For this experiment, the silicon die has $16mm \times 16mm \times 0.5mm$ dimensions. The thermal interface material (TIM) layer has the same size as the die and is 0.1mm thick. The TIM conductivity is $7.5W/(m \cdot K)$ as in [1]⁹. The results for a worse TIM with thermal conductivity of $1.33W/(m \cdot K)$ are not shown in this paper due to limited space. They are available in the extended technical report of this paper [9].

We also found that modeling heat spreader and heat sink separately as in HotSpot is more accurate than collapsing the two together as in [1]. Although the difference is not significant when both of them are made of copper, HotSpot is certainly more flexible to model the cases where heat spreader and heat sink are made of different materials. However, in order to closely repeat the experiments in [1], in the experiments presented here, the spreader and heat sink are artificially collapsed into one piece of copper with the size of $60mm \times 60mm \times 7.9mm$ as in [1]. The heat transfer coefficient at the top surface is $2777.7W/(m^2 \cdot K)$, which is equivalent to a single lumped convection thermal resistance of $0.1K/W$ to be consistent with HotSpot3.1. The floorplan is one that is similar to that of EV6. We slightly modify the coordinates of the functional blocks for alignment so that it is easier to build the model in ANSYS. We use the same modified EV6 floorplan for HotSpot, ANSYS and FreeFEM3d in this experiment. The floorplan is shown in Fig. 3.

Fig. 4(a) shows the temperature estimations from ANSYS, FreeFEM3d (FF3d), HotSpot3.1 and HotSpot4.0. To better illustrate the absolute errors of the HotSpot block model, in Fig. 4(b), we use ANSYS temperatures as the references and

⁸For example, when a layer is divided into two sub-layers, i.e. approximating the distributed system by a two-lumped RC ladder, SPICE simulation shows that a ~ 0.7 factor is needed to match the thermal time constants. Similarly, a ~ 0.8 factor is found for three-lumped RC ladder approximation.

⁹In HotSpot3.1, the default TIM thermal conductivity is $k_{TIM} = 1.33W/(m \cdot K)$ and its thickness is $75\mu m$. Notice that TIM with a $75\mu m$ thickness is really an extreme case. In state-of-the-art package, TIM is much thinner and usually has better thermal conductivity

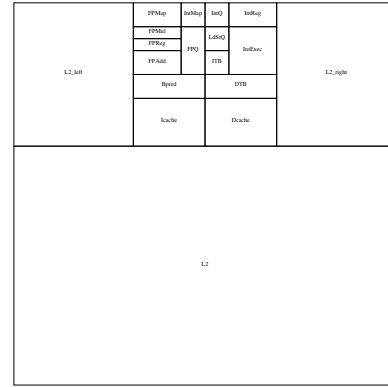


Figure 3. EV6 floorplan.

plot the errors of HotSpot4.0, HotSpot3.1 and FreeFEM3d (FF3d) with respect to ANSYS.

There are several observations from Fig. 4:

1. HotSpot 4.0 in general has fewer errors than HotSpot 3.1.
2. There are noticeable differences between ANSYS and FreeFEM3d (FF3d) as well, both are detailed finite-element models.
3. For the case of good TIM, HotSpot is even closer to ANSYS than FreeFEM3d! What is more, even HotSpot 3.1 does provide reasonably accurate temperature estimations. Since the package configuration with good TIM represents a realistic package for modern high-performance microprocessors, *we can see that the original HotSpot 3.1 block model is already quite accurate.*
4. For the case of worse TIM (not shown here, see [9]), HotSpot predicts hotter temperatures than both ANSYS and FreeFEM3d in most cases, but the percentage errors for hot units, e.g. BPred and IntReg, are 3.05% and 2.56%, respectively. Overall worst-case percentage error with worse TIM is 11.96% for I-Cache, which is a relatively cool unit.

The improved accuracy is achieved by eliminating the isotherm nodes in package and dividing high-aspect-ratio blocks into sub-blocks with unit aspect ratios. (see bullets 1, 2 and 3 in Section 2.1.1).

3.1.2 Square Source

A better experiment that helps to evaluate and explain the HotSpot block model steady-state errors is to sweep the heat source size with the same power density, as was done in [1]. Here, we also repeat this experiment and demonstrate the improved accuracy of HotSpot4.0.

In this experiment, the silicon chip has a size of $21mm \times 21mm \times 0.5mm$, and the dimensions of other package components are the same as Section 3.1.1. The center heat source size changes from 1mm to 19mm. The applied power density to the center block is set to a constant value of $1.66W/mm^2$, as in [1]. Fig. 5(a) shows a floorplan with a $1mm$ square heat source together with its high aspect ratio neighbor

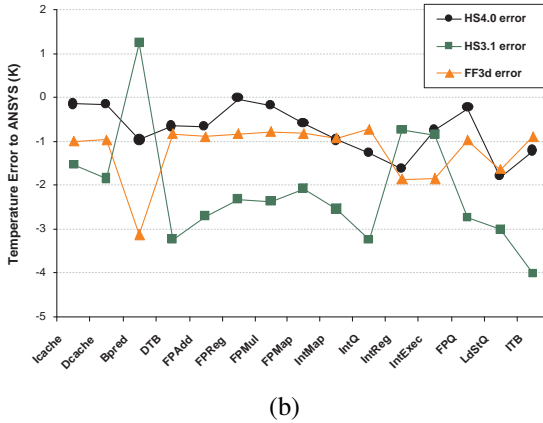
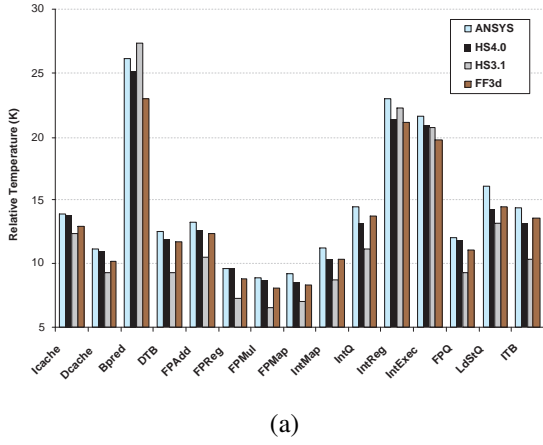


Figure 4. (a) EV6 block relative temperatures with good thermal interface material. (b) EV6 block relative temperature errors with respect to ANSYS, with good thermal interface material ($k_{TIM} = 7.5W/(m \cdot K)$).

blocks. Fig. 5(b) shows the same floorplan in which the high aspect ratio blocks are divided into square sub-blocks.

Fig. 6 shows the comparisons among the HotSpot 3.1, HotSpot 4.0, ANSYS and FreeFEM3d for different heat source sizes. We also plot the HotSpot 3.1 results with unity-aspect-ratio (sub)blocks (HS3.1-AR) to isolate the effect of each individual aforementioned modifications (i.e. unity aspect ratio and non-isothermal boundary condition). As can be seen, the HotSpot 4.0 block model is much more accurate than the earlier HotSpot 3.1 block model.

For smaller heat source size (1mm to 5mm), the significant error of HotSpot 3.1 is caused by the extreme aspect ratio (10:1) of the four long and skinny blocks that are adjacent to the center small heat source block. In HotSpot 4.0, these long and skinny blocks are divided into 10 sub-blocks with aspect ratios of 1:1, thus the accuracy is greatly improved (see left part of the “HS3.1 AR” curve for small heat source sizes).

For larger heat source size (e.g., 19mm, which has 600W of power!), the significant error of HotSpot 3.1 is caused by the fact that the upper surfaces of the heat spreader and the heat sink are no longer close to being isothermal, so approximating them with single nodes yields significant errors. In HotSpot 4.0, the isothermal nodes are removed. Instead, we model the heat sink

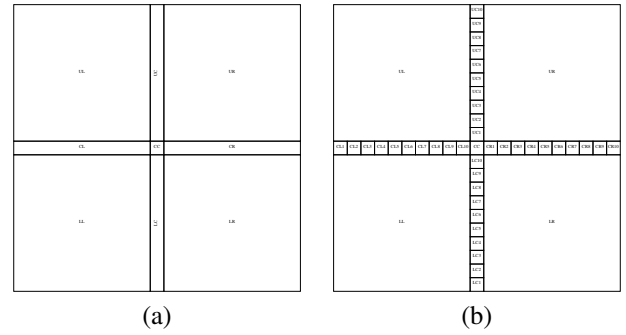


Figure 5. (a) Floorplan with 1mm center square heat source dissipating 1.66W. Notice the neighboring high aspect ratio blocks. (b) The neighboring high aspect ratio blocks are divided into square sub-blocks.

at the same level of detail as the silicon die and use a constant heat transfer coefficient ($h = 2777.7W/(m^2 \cdot K)$) for each sub-area of the heat sink-ambient interface, resulting in the same convective boundary condition as that in [1]. This significantly improves the accuracy for large-size heat sources (see the significant improvement for larger heat source sizes from “HS3.1 AR” to “HS4.0”). More results of this experiment with a worse TIM material can be found in [9].

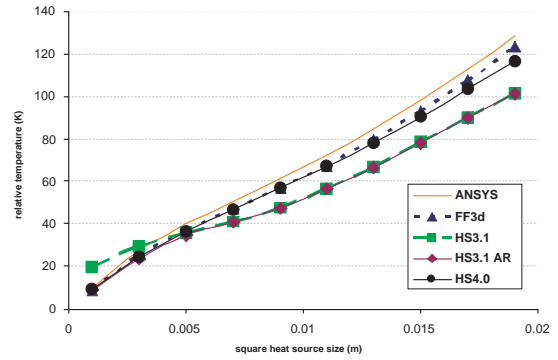


Figure 6. Center temperature for different heat source sizes, with good thermal interface material ($k_{TIM} = 7.5W/(m \cdot K)$), power density is $1.66W/mm^2$.

Here, again, by eliminating the isotherm nodes in package and dividing high-aspect-ratio blocks into sub-blocks with unit aspect ratios, HotSpot block model greatly improves the inaccuracies mentioned in [1] (see bullets 1, 2 and 3 in Section 2.1.1).

3.2 Transient Temperature

To validate the improved transient accuracy of HotSpot 4.0, we performed an experiment with power pulses of different time scales. More results on transient validation and comparison of HotSpot 4.0 can be seen in the technical report [9].

In Fig. 7, power pulses of $100\mu s$, 1ms and 10ms are sequentially applied to the Branch Predictor(Bpred) block in the EV6 floorplan with uniform power density of $2W/mm^2$ to verify HotSpot 4.0’s accuracy at different time scales. Notice the time axis is in log scale. We compare HotSpot 4.0 and HotSpot 3.1

results with ANSYS. As can be seen, HotSpot 4.0 significantly improves transient accuracy for all time scales under this high-aspect-ratio and high-power-density extreme case.

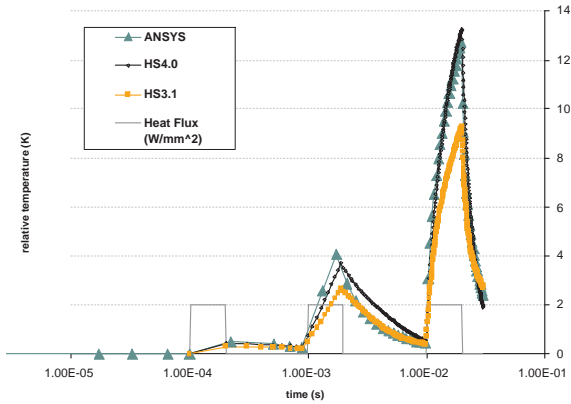


Figure 7. Transient temperature response for different power pulse widths applied to the branch predictor of EV6. Power density is $2\text{W}/\text{mm}^2$ ($k_{TIM} = 7.5\text{W}/(\text{m}\cdot\text{K})$).

We can see that in addition to eliminating the isotherm nodes in package and dividing high-aspect-ratio blocks into sub-blocks with unit aspect ratios, HotSpot block model’s transient accuracy is also improved by using a 0.5 constant scaling factor to approximate the thermal time constant of the distributive nature of transient temperature evolution. (cf. bullets 4 and 5 in Section 2.1.1).

Based on the above steady-state and transient experiments and comparisons among HotSpot block model, ANSYS and FreeFEM3d, we can see that HotSpot, especially the improved HotSpot 4.0 model, is accurate as a compact thermal model for architecture-level and other early-stage design levels. The small inaccuracies come from the fact that HotSpot trades off accuracy to achieve greater model compactness. The lumped nature of the HotSpot block model determines the fact that it cannot achieve the same level of accuracy as ANSYS or FreeFEM3d. But with small modifications, HotSpot block model’s accuracy can be greatly improved. Such improvements include (1) making the block lateral aspect ratio closer to unity; (2) applying a more realistic convective boundary condition by eliminating the isotherm nodes and using heat transfer coefficient; and (3) using proper scaling factors for the lumped thermal capacitors to approximate the distributive nature of transient heat transfer.

4 Thermal Modeling Granularity

Historically, in thermal analysis for early designs, a single temperature was used for the entire die. Similarly, in the HotSpot block model, the center temperature of a block is used to represent the entire block, while ignoring the temperature variations at finer granularities within the block. Whether this simplification is legitimate or not depends on several factors—the design level one works at, the available granularity of power estimation, the desired design complexity and accuracy trade-off, etc. For example, a temperature model at the transistor level is obviously less useful for designers who can only estimate power at the architecture level. Even if this designer has

access to transistor level power numbers, performing thermal simulations for the entire system at the transistor level results in prohibitive computation overhead.

However, whenever a more detailed estimation of power distribution within a functional block is available, it is still important to know whether ignoring localized heating within the block would miss potential local within-block hot spots that impact the reliability and performance of the entire design. In this section, we propose an analytical approach to find the relationship between the available power modeling granularity and the peak temperature rise at that granularity, i.e., the relationship between the size of a uniform heat source and its peak temperature. This analysis provides a straightforward way for researchers to decide whether it is necessary to model higher-resolution localized heating. It can also help design engineers to find the right granularity at which thermal monitoring and thermal management techniques are deployed, without missing hot spots on one hand or over-engineering to an unnecessarily fine granularity on the other. *The prerequisite, however, is that the finer granularity power distribution is available.* On the other hand, if the finer granularity power is not available, the best meaningful thing one can do is to do thermal modeling at the current design level.

There are a few existing works on thermal-related modeling granularity analysis. Etesam-Yazdani and Hamann et al. [10] investigated the thermal and power granularity issue by experimentally finding the relationship between the size of heat source and the peak temperature, and providing some guidelines for choosing the proper granularity. Our previous work [5] presented a preliminary analytical solution to find the proper thermal modeling granularity. In this paper, we propose a more rigorous analytical approach based on *the analogy between temporal frequency domain electrical circuit analysis and the spatial frequency domain thermal circuit analysis*. It can be easily understood and used by computer scientists and electrical engineers without digging into full details of the underlying heat transfer theories. This approach also theoretically explains the results reported by detailed finite-element simulations in [10, 5].

4.1 Theories

In mathematics, physics and image processing, *spatial* frequency is an attribute of any quantity that is periodic in space. It is a measure of how often that quantity is repeated per unit distance (e.g. per meter). It is defined as

$$f_s = \frac{1}{\lambda} \quad (3)$$

where f_s denotes the spatial frequency, λ is the period or wavelength of the repeating pattern.¹⁰

For the purpose of illustration, we first show a traditional temporal frequency-domain analysis for a first-order electrical RC circuit, we then utilize the analogy between the temporal frequency (in s^{-1} or Hertz) and the spatial frequency (in m^{-1}) to extend the analysis from time to space as well as from electrical domain to thermal domain.

For an electrical capacitor, C , assume the voltage drop between its two terminals is a sinusoidal form with frequency ω ,

¹⁰In particular, for our case of thermal granularity analysis in the spatial frequency domain, if the duty factor is 0.5, we can view λ as twice the size of heat source.

that is, $V_c(t) = V_0 \cos(\omega t + \phi)$, or in the exponential form, $V_c = V_0 e^{j(\omega t + \phi)}$. From the circuit theory, the current flow through the capacitor $I_c(t)$ is

$$I_c = C \frac{dV_c}{dt} = C \frac{d}{dt} V_0 e^{j(\omega t + \phi)} = j\omega C \cdot V_c \quad (4)$$

Thus, the *electrical* impedance of the capacitor is

$$Z_C = \frac{V_c}{I_c} = \frac{1}{j\omega C} \quad (5)$$

Now, consider the electrical circuit in Fig. 8, which has a resistor R , a capacitor C and a sinusoidal voltage source $V_s(t) = V_0 \cos(\omega t + \phi)$. We know that this circuit is a low-pass filter, that is, the voltage drop across the capacitor tracks the input voltage $V_s(t)$ at low frequency, and is increasingly attenuated at higher frequency. In other words, the equivalent impedance of this circuit is $Z_{eq} = Z_R || Z_C = R || (\frac{1}{j\omega C})$, with $Z_{eq} = R$ at DC, and approaching zero at high frequencies, thus comes the term “low-pass filter”. The resistor R determines the “DC” component of the output voltage, whereas the capacitor determines the “AC” component.

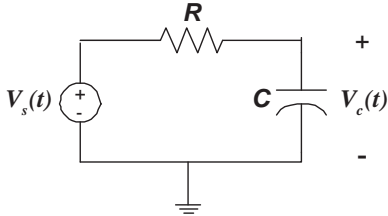


Figure 8. A first-order electrical RC circuit.

In space, there is also this “low-pass filtering” effect for temperature distribution. Here, we extend the temporal frequency analysis to the one-dimensional spatial frequency domain. Consider a sinusoidal heat flux (power density) of $q(x)$, which causes a sinusoidal temperature distribution

$$T(x) = T_0 \cos(\omega_s x + \phi) = T_0 e^{j(\omega_s x + \phi)} \quad (6)$$

where $\omega_s = 2\pi/\lambda$ is the spatial radian frequency (the subscript s means “spatial”), and x is the position in the 1-D space. The governing equation of heat transfer is Fourier’s Law

$$q(x) = k \frac{dT(x)}{dx} = k \frac{d}{dx} T_0 e^{j(\omega_s x + \phi_s)} = j\omega_s k T(x) \quad (7)$$

where k is the thermal conductivity. The minus “-” sign in Fourier’s Law goes away if we define $dT(x)$ as the temperature decrease (high temperature minus low temperature). Notice the similarity between Eq. (7) and Eq. (4). This leads us to some quantity analogous to the electrical capacitor in the spatial domain for heat transfer. We call it *thermal spatial capacitive impedance*, and write it as

$$Z_{C_s} = \frac{T}{q} = \frac{1}{j\omega_s k} = \frac{1}{j\omega_s C_s} \quad (8)$$

where C_s is defined as *thermal spatial capacitance* (notice that C_s is completely unrelated to the thermal capacitance C_{th} that we defined earlier in this paper that determines the *transient* heat transfer), and Z_{C_s} is the “thermal spatial capacitive

impedance”. The subscript “s” denotes the spatial nature of these definitions. The unit of both C_s and Z_{C_s} is $m^2 K/W$, which is different from the unit of the thermal resistance we used earlier in this paper (in K/W). This is legitimate because we use heat flux, i.e. power density (in W/m^2), instead of power (in W). In other words, the thermal impedance and resistance in this section is defined as the temperature drop divided by the power density, not by power.

Eq. (8) is used when there is an AC component, with spatial frequency ω_s , in the applied heat flux. In the case where there is only DC heat flux, Fourier’s Law leads to the traditional definition of thermal resistance

$$Z_{R_s} = \frac{t_{eq}}{k} \quad (9)$$

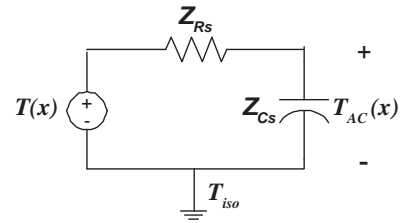


Figure 9. The Thevenin equivalent first-order thermal spatial “ RC ” circuit.

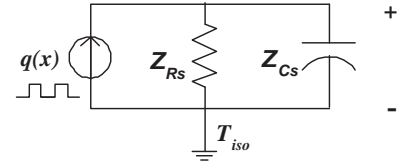


Figure 10. The Norton equivalent first-order thermal spatial “ RC ” circuit.

where t_{eq} is the distance from the active silicon surface to the isotherm surface in the package. Also note that this DC spatial thermal impedance also has the unit of $m^2 K/W$, which is consistent with the unit of the AC spatial thermal impedance Z_{C_s} . From the above derivation, naturally we can reach a first-order spatial thermal “ $R_s C_s$ ” circuit as shown in Fig. 9. To make it more comprehensible, Fig. 10 shows a more intuitive Norton equivalent circuit of Fig. 9. The heat flux generated by the active silicon layer is written as $q(x)$, which models the non-uniform distribution of power density across the chip. The DC component in the spatial temperature distribution is determined by Z_{R_s} , whereas the AC component is determined by Z_{C_s} . In addition, the total equivalent thermal spatial impedance is

$$Z_{eq_s} = Z_{R_s} || Z_{C_s} \quad (10)$$

If we plot the Bode plot of Z_{eq_s} with respect to the spatial frequency ω_s in Fig. 11, we can see that for low spatial frequencies (power sources with large dimensions), the thermal impedance is close to the DC component, that is the lumped $R_{th} = t_{eq}/(kA)$ as we used in earlier sections (A is the corresponding vertical heat conduction area). But for high spatial frequencies (power sources with small dimensions), the impedance

attenuates to smaller values due to the presence of the thermal spatial “capacitance”. This explains the spatial temperature filtering effect—structures with tiny dimensions have lower peak temperature comparing to their larger counterparts applied with the same power density. Fig. 12 illustrates an example of this low-pass filtering effect by comparing the HotSpot 4.0’s results of the temperature rise curves for different heat source sizes with power densities of $1.66\text{W}/\text{mm}^2$ and $5\text{W}/\text{mm}^2$. For example, from Fig. 12, a 5mm heat source with $5\text{W}/\text{mm}^2$ has the same temperature rise as a 18mm heat source with $1.66\text{W}/\text{mm}^2$. Notice that in the figure, the curves of different power densities do not cross. This is because for the same size, higher power density always results in higher temperature. Actually, once a curve is obtained at one power density, curves of other power densities can be easily derived by linear scaling. This is because the thermal RC network is linear as long as the thermal conductivity is not temperature dependent.

Intuitively, a tiny heat source even with a high power density does not significantly increase the total power dissipation that the package has to remove, thus temperature rise in the package is almost negligible. On the other hand, a large heat source with high power density results in significant rise in total power dissipation, which in turn leads to significant temperature rise at the heat sink and the heat spreader, hence the increase of average and peak silicon temperatures.

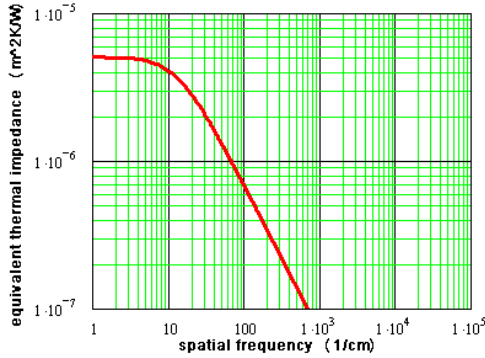


Figure 11. The thermal spatial “RC” circuit is low-pass filter in the spatial frequency domain.

In [10], Etessam-Yazdani and Hamann et al. also reached a similar spatial frequency domain low-pass filter characteristic curve for square wave power distributions by experimental simulations in finite-element tools. Notice that a $2/\pi$ scaling factor needs to be multiplied to the radian frequency in the above derivations to account for the peak temperature difference between a sinusoidal input power density pattern and a square-wave pattern in real designs.

Because the heat transfer in x and y lateral directions are orthogonal, which is determined by the 2-D form of Fourier’s Law, the above derivations can also be easily extended into two-dimensional space with similar results.

One limitation of the above analysis is that it takes into account the lateral spatial temperature gradient, but not the vertical gradient. A more accurate analysis would be using multiple R_sC_s ladders (i.e. dividing each layer vertically into multiple sub-layers), or ideally, distributed thermal spatial thermal R_sC_s circuit. Fig. 13 shows the comparison between the proposed

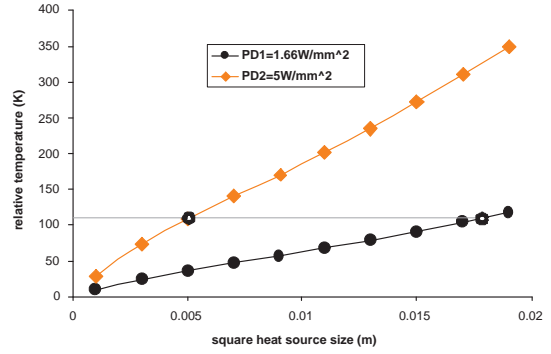


Figure 12. A HotSpot 4.0 example of the low-pass spatial temperature filtering effect. As the horizontal line indicates, smaller heat source (5mm) with a high power density ($5\text{W}/\text{mm}^2$, top curve) results in the same temperature rise as a large heat source (18mm) with a lower power density ($1.66\text{W}/\text{mm}^2$, bottom curve).

granularity analysis (3-ladder spatial R_sC_s circuit) and ANSYS simulations for different heat source sizes. Note that the spatial frequency and equivalent thermal impedance are both normalized.

As can be seen in Fig. 13, as long as the heat source size is about ten times greater than the isothermal thickness, the thermal resistance can be calculated using conventional $R_{th} = t/(kA)$. For smaller heat sources, the spatial low-pass filtering effect kicks in, and the effective thermal resistance is much less. This means that *tiny heat sources are not necessarily hot spots even with very high power densities*. For example, assuming the isotherm thickness is 4mm, for a heat source of 0.1mm size, we have a normalized spatial frequency of 40, which corresponds to $0.045 \times$ the peak resistance from Fig. 13, resulting in $0.045 \times$ peak temperature rise. In other words, if a large heat source leads to 100C temperature rise, this 0.1mm heat source with the same power density only contributes to 4.5C temperature rise. This also explains why some high power density tiny structures, such as clock buffers, do not always become local hot spots. It is obvious that the low-pass temperature filtering effect for the relationship between heat source size and peak temperature is strong.

4.2 Implications of the Granularity Analysis

With the above thermal granularity analysis, it is easy to know whether the spatial power fluctuation at a finer granularity would cause significant inaccuracy for temperature estimation at a coarser granularity, as long as one knows the heat source size, the local power density and the isothermal thickness of the package. Typical isothermal thickness in high-performance microprocessor packages is about a few millimeters to our knowledge. In comparison, most functional units at the micro-architecture level have similar or larger sizes. On the other hand, sub-unit blocks that cause non-uniform power distributions usually have sizes of a few hundred microns or less, such as individual register file entries, cache lines, small ALU modules, etc. Therefore, equivalent thermal resistances for those sub-units are usually 10 times or more smaller due

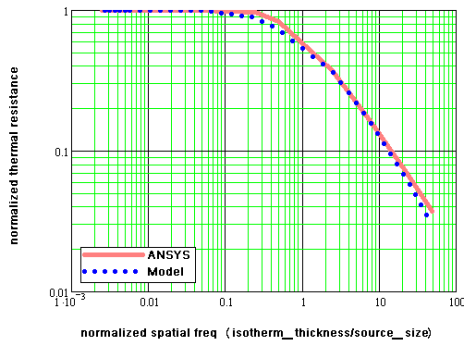


Figure 13. Comparison of 3-ladder thermal spatial RC model and ANSYS simulation for different heat source sizes.

to the spatial temperature filtering effect. As a more extreme example, Xiu et al [11] reported a finite-element simulation for a $15\mu\text{m} \times 300\mu\text{m}$ circuit macro (a clock buffer) dissipating $25\text{W}/\text{mm}^2$ (average chip power density is only $0.4\text{W}/\text{mm}^2$). The resultant peak steady-state temperature of this small macro is only 4.5 to 7.2C higher than the other part of the die that consumes average power for different boundary conditions. On the other hand, if such a high power density is applied to a large block with 4mm equivalent silicon thickness, we would see a $\sim 1000\text{C}$ steady-state temperature rise.

In addition, most sub-blocks are usually switched on for a short period. For example, a typical read access to an individual cache line only lasts for a few nanoseconds. In other words, the power is only dissipated on the cache line for a short time. Therefore, the resultant transient temperature evolution time (ns) is much shorter than the thermal time constant (μs or ms). This means that temperature already starts to drop before it reaches its potential steady-state value (in which case the power is assumed to be dissipated continuously all the time). We call this phenomenon *temporal temperature filtering effect*, in contrast to the *spatial temperature filtering effect* we have just described.

The combined effect of spatial and temporal temperature filtering effects is that tiny sub-units will not usually cause serious local hot spots, unless their power density is much higher (10 times or more) than the overall power density of the entire block, and the high power density is applied constantly. Only if that happens, the thermal model needs to go down to a finer granularity to fully account for the localized heating within each block. (Remember that the prerequisite is that the finer granularity power estimations are available.)

5 Conclusions

In this paper, we first present improvements to the HotSpot 3.1 block model to make it accurate even under high lateral block-aspect ratios, high power density, and to better model realistic convective boundary conditions. The accuracy improvements of both steady-state and transient temperatures are confirmed by comparing with finite-element model in ANSYS, a well-established commercial package.

In addition, we also propose an analytical approach to determine the necessary spatial thermal modeling granularity to

make HotSpot more useful. This analysis derives the relationship between heat source size and peak temperature and finds that small heat source often leads to lower peak temperature for the same power density. The results can help guide researchers and engineers to the proper granularities at which to invest temperature-aware design efforts.

6 Acknowledgement

The authors would like to thank Damien Fetis and Pierre Michaud from IRISA/INRIA, France, for their generous help with FreeFEM3d and interesting discussions. This work is funded by NSF CRI grant CNS-0551630.

References

- [1] D. Fetis and P. Michaud. An evaluation of hotspot-3.0 block-based temperature model. In *Proc. 2006 WDDD, in conjunction with ISCA 2006*, June 2006.
- [2] K. Skadron, K. Sankaranarayanan, S. Velusamy, D. Tarjan, M. R. Stan, and W. Huang. Temperature-aware microarchitecture: Modeling and implementation. *ACM Transactions on Architecture and Code Optimization*, 1(1):94–125, March 2004.
- [3] D. Brooks, V. Tiwari, and M. Martonosi. Wattch: A framework for architectural-level power analysis and optimizations. In *Proc. Intl. Symp. of Computer Architecture (ISCA)*, pages 83–94, June 2000.
- [4] K. Skadron, M. R. Stan, W. Huang, S. Velusamy, K. Sankaranarayanan, and D. Tarjan. Temperature-aware microarchitecture. In *Proc. Intl. Symp. on Computer Architecture (ISCA)*, pages 2–13, June 2003.
- [5] W. Huang, S. Ghosh, S. Velusamy, K. Sankaranarayanan, K. Skadron, and M. R. Stan. Hotspot: A compact thermal modeling methodology for early-stage vlsi design. *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, 14(5):501–513, May 2006.
- [6] S. Velusamy, W. Huang, J. Lach, M. R. Stan, and K. Skadron. Monitoring temperature in fpga based socs. In *Proc. International Conference on Computer Design (ICCD)*, pages 634–637, October 2005.
- [7] S. Lee, S. Song, V. Au, and K. Moran. Constricting/spreading resistance model for electronics packaging. In *Proc. American-Japan Thermal Engineering Conf. (AJTEC)*, pages 199–206, March 1995.
- [8] K. Skadron, M. R. Stan, W. Huang, S. Velusamy, K. Sankaranarayanan, and D. Tarjan. Temperature-aware microarchitecture: Extended discussion and results. Technical Report CS-2003-08, University of Virginia, Computer Science Department, 2003.
- [9] W. Huang, K. Sankaranarayanan, R. J. Ribando, M. R. Stan, and K. Skadron. An improved hotspot block-based thermal model with granularity considerations (extended discussion). Technical Report CS-2007-07, University of Virginia, Computer Science Department, 2007.
- [10] K. Etessam-Yazdani, H. F. Hamann, and M. Asheghi. Impact of power granularity on chip thermal modeling. In *Proc. of 10th Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronics Systems (ITHERM)*, pages 666–670, June 2006.
- [11] K. Xiu and M. Ketchen. Thermal modeling of a small extreme power density macro on a high power density microprocessor chip in the presence of realistic packaging and interconnect structures. In *Proc. of Electronic Components and Technology Conference (ECTC)*, pages 918–923, 2004.