Abstract—The degraded thermal path of 3-D integrated circuits (3DICs) makes thermal analysis at the chip-scale an essential part of the design process. Performing an appropriate thermal analysis on such circuits requires a model with junction-level fidelity; however, the computational burden imposed by such a model is tremendous. In this paper, we present enhancements to two thermal modeling techniques for integrated circuits to make them applicable to 3DICs. First, we present a resistive mesh-based approach that improves on the fidelity of prior approaches by constructing a thermal model of the full structure of 3DICs, including the interconnect. Second, we introduce a method for dividing the thermal response caused by a heat load into a high fidelity “near response” and a lower fidelity “far response” in order to implement Power Blurring high definition (HD), a hierarchical thermal simulation approach based on Power Blurring that incorporates the resistive mesh-based models and allows for junction-level accuracy at the full-chip scale. The Power Blurring HD technique yields approximately three orders of magnitude of improvement in runtime for a three-tier 3DICs, however, there is a significant limitation: heat flow inhibiting the heat flow, which places tiers far from the heatsink at the overall thermal profile to become more pronounced. This causes the effect of detailed structures removed when silicon-on-insulator (SOI) technology is used on semiconductor devices include reduced threshold voltage, increased subthreshold leakage [2], and decreased mean time to failure due to electromigration [3]. Changes in device performance with temperature necessitate the consideration of on-chip temperature variations when attempting to obtain timing closure in digital circuits [4], [5].

For modern very large scale integration circuits, the exact calculation of the thermal profile is a daunting task. High-performance integrated circuits routinely contain millions of logic gates, memory cells, and other structures. Each of these structures has its own complex 3-D structure, including wires, vias, and transistors. The thermal properties of the materials used to fabricate these circuits vary highly. The silicon substrate, which typically acts as a heat spreader in 2-D chips, is removed when silicon-on-insulator (SOI) technology is used as part of a 3DIC. This causes the effect of detailed structures on the overall thermal profile to become more pronounced. Oxide layers such as those between the tiers of a 3DIC can inhibit the heat flow, which places tiers far from the heatsink at a thermal disadvantage [6]–[8]. Additionally, low-k dielectrics have a lower thermal conductivity [9], which further degrades thermal performance.

Many techniques have been previously presented for full-chip thermal analysis, with a representative sample being [4], [10]–[12]. When we consider applying these methods to SOI 3DICs, however, there is a significant limitation: heat flow between tiers occurs through a composite material of metal and insulator, rather than through the substrate. Finite element method (FEM) or finite difference method (FDM) solvers can be used to calculate an average thermal conductivity for regions of the chip, but such calculations require a memory-intensive matrix solution. To avoid this, the weighted
Average (parallel) model has been used by both Li et al. [11] and Zhan et al. [10] to estimate the thermal conductivity. This parallel model along with its counterpart, the series model [13], are the two fastest methods to evaluate average thermal conductivity. These two models also serve as bounds on the thermal conductivity. For a copper-oxide composite with a ratio of 20% copper to 80% oxide volume (typical for a 3DIC), the estimates computed by these approaches differ by nearly two orders of magnitude. This means that predicted temperatures may vary by as much as a factor of 100, depending on which method is used to calculate average conductivities. This paper addresses the inaccuracies caused by using the series or the parallel model by performing the detailed matrix solution needed to determine the heat flow through a metal-oxide composite material. It is demonstrated that an average thermal conductivity model for the composite works well to predict temperatures at points far away from the transistor, and that this average conductivity does not vary much across the chip. Near each transistor, the response calculated through the matrix solution provides a good estimate of temperatures. This paper also extends the Power Blurring technique introduced by Kemper et al. [12] to collect these “near” and “far” estimates into a full-chip estimate.

The remainder of this paper is organized as follows. Section II gives an overview of thermal simulation approaches. Section III introduces the 3-D technology used in this paper. Section IV describes the thermal network extractor. Section V describes the Power Blurring HD approach. Section VI uses Power Blurring HD to analyze an FFT processor for a laboratory’s 3-D integration process, which was used to fabricate chip simulation with junction-scale accuracy is intractable for these approaches. Approaches based on Green’s function [10], [15]–[18] avoid the requirement to mesh the entire circuit and instead determine the thermal response due to a single point source, with the full-chip response obtained through the use of superposition. These techniques are faster than traditional FEM/FDM solutions, but rely on homogenizing assumptions to simplify the problem. Inherent to all Green’s function-based approaches is the assumption that the thermal conductivity on each layer is homogeneous. In [16], square geometries are considered and look-up tables are used to speed up the calculation; however, the approach is still cost prohibitive for full-chip simulation when there are numerous sources. In [18], an analytical form for multilayer Green’s function analysis is derived that is capable of considering sidewall boundary conditions; however, this approach’s inability to handle variations in the thermal conductivity of individual layers makes it unsuitable for analyzing SOI 3DICs.

In [11], an efficient multigrid-based approach is introduced that avoids the explicit construction of the matrix problem. Small volumes are modeled using the weighted average of the thermal conductivities of the materials enclosed, leaving more accurate modeling of small areas as an open research question. Power Blurring [12], [19]–[25] is a superposition-based approach where an “impulse response” describing the thermal response to an impulse power source is convolved with a power map to determine the full-chip thermal profile (see Fig. 1). Power Blurring shares similarities with Green’s function-based approaches; however, unlike these approaches it is not constrained to using a homogeneous thermal conductivity for each layer. Methods for determining response masks are varied and can include an FEM/FDM simulation or thermal measurement, such as with an infrared camera.

In our prior work, we have shown that simplifying the thermal model of a circuit by using homogeneous thermal conductivities to describe model layers can significantly underpredict channel temperatures in SOI 3DICs, particularly for devices that are far from the heatsink [26]. With the exception of [11], none of the presented approaches are suitable for handling SOI 3DICs. Park et al. [23] previously proposed 3-D extensions to Power Blurring and while these extensions have been shown to be suitable for non-SOI 3DICs, the modeling fidelity is not sufficient for SOI 3DICs. The Power Blurring method is, however, general enough that with the proper modifications it can be made suitable for SOI 3DICs.

### III. MIT Lincoln Laboratory’s 3-D Process

In this section, we briefly describe the MIT Lincoln Laboratory’s 3-D integration process, which was used to fabricate...
Fig. 2. MIT Lincoln Laboratory’s 3-D process stackup [27]. The handle silicon for the bottom tier remains intact. The handles from the middle and upper tiers are thinned, and then wafer-bonded upside down. TSVs are used to connect signals between the tiers. Note: layer thicknesses are not to scale.

TABLE I

<table>
<thead>
<tr>
<th>Thermal Units</th>
<th>Electrical Units</th>
</tr>
</thead>
<tbody>
<tr>
<td>Temperature (K)</td>
<td>Voltage (V)</td>
</tr>
<tr>
<td>Power (W)</td>
<td>Current (A)</td>
</tr>
<tr>
<td>Thermal resistance (K/W)</td>
<td>Electrical resistance (V/A)</td>
</tr>
<tr>
<td>Thermal capacitance [(W s)/K]</td>
<td>Electrical capacitance [(A s)/V]</td>
</tr>
</tbody>
</table>

IV. RESISTIVE MESH-BASED MODELING

In this section, we present the resistive mesh-based modeling technique that is used in this paper to generate response masks for Power Blurring HD. This modeling is also used to estimate the runtime and memory requirements of finite difference models for full-chip junction-level thermal analysis. This modeling is similar to that used by Cheng et al. [4] and Skadron et al. [28], but performed at a significantly higher resolution. The well-known correspondence between electrical and thermal variables (also used by [4] and [28], and shown in Table I) was used to solve the thermal equations by substituting the thermal equations with the corresponding electrical equations.

This modeling approach discretizes the full 3-D structure of the circuit into volumetric elements that are represented thermally with six resistors (6R), as shown in Fig. 3. Material information for each element is gathered using a Python script that retrieves layout information directly from an OpenAccess database and combines it with process parameters. The resistances used to model each block are calculated based on the weighted average of the thermal conductivities of the materials that comprise that block. The mesh of six resistor nodes is then represented by an equation that takes the form

\[ M\mathbf{t} = \mathbf{p} \]  

where \( M \) is the modified nodal admittance matrix, \( \mathbf{t} \) is the temperature vector, and \( \mathbf{p} \) is the external power and temperature source vector. Equation (2) is then solved using the conjugate residual method on sparse matrices with the PETSc library [29]. Due to the large size of the problem, it is vital that sparse matrix techniques be employed. In order to ensure accurate modeling, the grid size has to be sufficiently small and the location of the power dissipation has to be properly aligned to the grid as shown in Fig. 4(a). If the grid is not properly aligned, the area of the channel can no longer be clearly described by the grid squares, which results in two problems. First, the location of the power source itself spreads out. This artificially lowers the power density and can cause the simulated temperature in the channel to be lower than the actual temperature. Second, if the grid squares are misaligned the model will differ from physical reality. For example, forcing the model to dissipate power in the silicon dioxide will result in artificially high simulated temperatures because silicon dioxide has a much lower thermal connectivity than the channel.

Prior work on resistive mesh models was unable to allocate individual grid points to logic gates, nor consider transistor-level granularity [15]. This paper individually meshes each physical layer of the circuit in the z-direction, with additional meshing for thick layers, as shown in Fig. 5. The resistive mesh uses a weighted average to determine the thermal conductivity of each block; however, when the grid size nears the minimum transistor length in a process, as is the case in
At this point the model is less of a weighted average model and much closer to an exact model of the materials. Fig. 6 shows the simulated temperature rise for a cross-section of a 0.2 μm × 2 μm wide transistor for various lateral modeling resolutions. The transistor’s channel temperature rise is significantly underpredicted when the modeling granularity in the x and y-dimensions is not at least as small as the smallest transistor channel dimension, with subsequent reduction having virtually no effect on the predicted temperature. Due to the resolutions required to accurately model the structure of the chip, resistive mesh-based models are limited in the area that they can model while maintaining their accuracy. The problem size is primarily limited by available system memory, with 30 million nodes being near the limit for a machine with 32 GB of RAM. Once the memory limit is reached, a divide and conquer approach must be used to handle larger problems.

V. POWER BLURRING HD

Power Blurring HD is based on the Power Blurring technique introduced by Kemper et al. [12] and is applicable to both planar ICs and 3DCs. Power Blurring HD can be used on both SOI and non-SOI circuits; however, due to the heat spreading in the substrate of non-SOI circuits we expect that such circuits could be adequately modeled by existing lower fidelity approaches, such as [12] or [23].

When compared to a traditional finite difference approach, such as the resistive mesh-based approach presented in the previous section, the Power Blurring HD approach offers several orders of magnitude of improvement in both the memory requirements and runtime for a full-chip simulation with junction-scale accuracy.

In this paper, the Power Blurring technique introduced by Kemper et al. [12] has been extended to:

1) consider complex structures where transistors are located in a sea of metal and a substrate is not always readily available for heat spreading;
2) consider chips designed in silicon-on-insulator processes through the inclusion of high fidelity material information for accurate simulation at the transistor scale;
3) handle chips whose size makes it intractable to calculate the full-chip response of a single heat source.

Power Blurring HD relies on pre-calculated “response masks” that describe the thermal response of the circuit structure given a pre-determined heat load. Power Blurring HD is performed by executing the following steps.

1) A set of heat loads are selected and the response masks for those heat loads are calculated.
2) A power map is generated for each response mask to describe the location and power dissipation of heat loads to be modeled with that response mask.
3) Response masks and power maps are convolved to determine the full-chip response.

A. Chip Characterization

Response masks can be generated using a variety of methods. The fundamental technique is to apply a known heat load to a single transistor in the design and observe the thermal response. This could be done with either simulation or measurement, where appropriate. Ideally, the full-chip response
Fig. 7. Temperature profile on the active layer of Tier B for a resistive mesh-based simulation at 0.1 μm of a transistor on Tier B. The temperature response depends highly on the orientation of the materials, with the vertical line of high temperatures lining up with the transistor channel where power was applied.

to the heat load would be calculated; however, this is often infeasible due to the size of the model that would be needed to accurately capture the gradients in the thermal response near the transistor with the applied heat load. By modeling only part of the chip, a tradeoff is made in terms of accurately determining the channel temperature of the single transistor at the expense of losing information about that transistor’s impact on locations far from it. This is an appropriate simplification since the gradient of the thermal profile due to a single heat load decreases as one moves further from the transistor with the applied heat load.

Separate response masks are generated for every homogeneous environment where transistors are present. The presence of fill shapes helps to homogenize the thermal properties of different regions of the chip. This means that even if one area of the chip has slightly higher density routing than another area of the chip, differences between the response masks that would be found for each region will be low. The actual differences between response masks will be presented in Section VI-B.

In the original Power Blurring approach, the response to a single point was used; however, in SOI circuits the orientation of a given transistor on an active island has a significant impact on the thermal response. Fig. 7 shows the temperature profile of a 2 μm transistor on Tier B. The hottest vertical line occurs directly over the transistor channel. While arbitrary power dissipation profiles can be handled through the use of response masks for point sources, this would be computationally expensive in SOI processes. Numerous point sources would be needed for a single transistor channel heat load to accurately model the orientation of the channel on the active island and the localized differences in heat spreading from each point in the channel [30]. For this reason, it is recommended that response masks be calculated at the transistor level or higher with Power Blurring HD. Individual transistors or entire standard cells are suitable candidates for response mask generation with a single entry in the power map for each transistor or standard cell. The downside to using standard cells is that the computational burden to characterize the chip will be high when numerous cells are used. In the future, this characterization information could be pre-computed and provided along with standard cell libraries, in the same manner that timing characterization currently is. Transistor level responses have been used for the chip characterization shown in the later sections of this paper. When using transistor level response masks, at least one response mask should be generated for every size of transistor in the design. Some approximation of the transistor size is acceptable and results in a tradeoff between the time needed to characterize the circuit and the accuracy of the transistor temperatures. The time needed to characterize the chip increases linearly with the number of response masks, assuming that the sizes (areas) of the response masks are identical.

In 3DICs, multiple layers of transistors are stacked on top of each other. Therefore, when calculating the temperature at a specific transistor, we must consider the transistors stacked above and below it. This is accomplished by including multiple layers of information in each response mask. The full thermal response on a specific layer is calculated by convolving all transistors with an appropriate response mask and then summing the effects of all of the response masks on the layer of interest.

B. Near/Far Effect

A series of simulations were used to examine the effect of using simulation areas smaller than the full chip. A single transistor from a clock buffer in the SAR processor that will be examined in Section VI was chosen for this comparison. The simulation areas were selected by starting at the boundary of the transistor channel and growing the area on all four sides by 8 μm, 16 μm, 32 μm, and 48 μm. Fig. 8 shows these
Fig. 9. Slices of the response mask for a transistor are shown based on simulations with varying amounts of material around the transistor being considered. (a) Responses as originally calculated without correction. (b) Responses after the offsets are removed and the far responses are added.

A graph of the responses on Tier B due to the heat generated by a transistor on Tier B is shown in Fig. 9(a). As the simulated area is decreased, it is evident that the offset of the entire simulation increases. Summing all values in the layer’s response matrices resulted in a nearly identical value (within 0.2%) for all of the simulations considered. The offset seen in these simulations is an artifact of the finite simulation area that was used. Finite simulations must be used because solving for the full-chip response with this type of model would be intractable.

Removing the offset by itself is not a complete solution, as the heat represented by the removed values would have spread out into the surrounding chip area if the area would have been present in the simulation. This can be approximated by dividing each response into a near and far response. The near response represents the area encompassed by the simulation done for the response mask. As illustrated in Fig. 10, the response mask determined by a bounded chip area simulation has both a sharp cutoff between the near and the far response and also overestimates the edge of the near response. This edge can be compensated by first removing the offset of the response mask. Convolution at the chip-scale is performed using the responses once the offsets have been removed. After the convolution is complete, the offsets are added.

For 3DICs, the offset is removed from each tier individually. After performing convolutions separately for each tier, the respective offsets are added back using the same ratio as for 2-D circuits.

Quantitatively speaking, the “near response” is the area around a heat source that is modeled by the response mask. The “far response” is the remaining area of the chip. In more qualitative terms, “far” can be seen as the point beyond which an average thermal conductivity model yields little error. The optimal size for the near response for a particular transistor can be determined with the following approach.

1) Calculate several sizes of response masks (“near” sizes) for the transistor of interest (Fig. 8).
2) Take slices of these response masks through the transistor (Fig. 9(a)).
3) Correct the slices by removing the offsets and adding the far responses (Fig. 9(b)).
4) Examine the slices to determine the minimum response mask size (“near” size) for which the slices begin to converge.

It is important to note that the minimum response mask size that is needed for convergence will be dependent on the technology and the layout shapes near the transistor.

C. Power Blurring HD Algorithm

The full algorithm for Power Blurring HD is shown in Fig. 11. This algorithm is suitable for calculating the thermal profile of both traditional (planar) ICs and 3DICs by taking into account the vertical coupling between heat sources on different tiers. In the algorithm, area(R) refers to the quantity $M \times N$ where $R$ is an $M \times N$ matrix and $R(L)$ refers to the temperature profile on layer $L$ in response mask $R$. The response mask characterization steps and chip-scale thermal profile calculation steps have been combined for clarity; however, they can be performed separately.

In the algorithm, a “homogeneous area” is defined as a group of transistors that can be accurately modeled with the same response mask. The algorithm initially relies on the designer to use their intuition to group transistors into homogeneous areas. These assumptions can then be tested.
define a group of transistors that can be accurately modeled with the same response mask.

by calculating several trial response masks. If two transistors at differing locations generate significantly different response masks, they should be considered to be in different homogeneous areas and should be modeled with separate response masks.

The algorithm will now be presented mathematically. The near response on Layer X due to transistors on Layer Y is defined as follows:

\[
N_{XY} = \sum_{i,j} (R_{XY} - \min(R_{XY})) \cdot P_Y^i
\]

where \( S \) is the set of response masks that have been calculated to characterize a given design.

\( R_{XY} \) denotes the response mask matrix for Layer Y given a heat load applied on Layer X that is characterized by response mask \( s \). The response mask matrix size is denoted by \( N \times M \) and the full chip size is denoted by \( M \times N \). \( P_Y^i \) is the power map of devices that use response mask \( s \) on Layer Y.

The far response on Layer X due to transistors on Layer Y is defined as follows:

\[
F_{XY} = \sum_{i,j} P_Y^i \cdot (\min(R_{XY}) - \frac{E_{s(i,j)}}{MN})
\]

where \( P_Y^i \) is the total power dissipated in devices that use response \( s \) on Layer Y.

\( E_{s(i,j)} \) is the power dissipated in a single clock buffer transistor on Tier B in the full design.

The approximated full temperature profile on Layer X is then the sum of the near and far responses

\[
T_X = \sum_{i,j} (N_{XY} + F_{XY})
\]

where \( L \) is the set of layers with heat dissipating devices on them.

A. Validation

In this section, Power Blurring HD is validated against the resistive mesh-based modeling. This was done by cutting a small region from the SAR. The size of the region was chosen so that it was close to the upper limit of the area that could be modeled with our hardware resources at a 0.1 \( \mu \)m resolution using the resistive mesh-based model. The drawn length of all transistors in this design is 0.2 \( \mu \)m.

Power was applied to 12 transistors across three cells of interest in this region: an or-and-invert gate on Tier A, a clock buffer on Tier B, and an and-or-invert gate on Tier C. The cells are located in a sea of other cells and a sea of metal interconnect. Metal fill structures were placed to meet the density requirements of the process and are fully modeled by both approaches. Each of the transistors was assigned a power value equal to that dissipated in a single clock buffer transistor on Tier B in the full design.

A total of six response masks were used for the Power Blurring HD simulation of this design. 2 \( \mu \)m and 4 \( \mu \)m wide transistors were modeled on Tiers A, B, and C. Each simulation was done at a resolution of 0.1 \( \mu \)m. For each response mask, profiles were recorded for the active layers on each tier as well as the top surface of the chip. This allows the full-chip simulation to account for the vertical coupling between the tiers. The 2 \( \mu \)m response masks are shown in Figs. 13–15.

Fig. 12. Micrograph of the FFT processor for SAR which was fabricated in the MIT Lincoln Laboratory 3-D process.
Fig. 13. 3-tier response mask for the MITLL Process calculated with a resistive mesh-based model for a heat load on a 2 \( \mu \text{m} \) wide transistor on Tier A. \( x \)- and \( y \)-axes are distance in \( \mu \text{m} \), \( z \)-axis is temperature rise over the heatsink in K. (a) Tier A. (b) Tier B. (c) Tier C. (d) Surface.

Fig. 14. 3-tier response mask for the MITLL Process calculated with a resistive mesh-based model for a heat load on a 2 \( \mu \text{m} \) wide transistor on Tier B. \( x \)- and \( y \)-axes are distance in \( \mu \text{m} \), \( z \)-axis is temperature rise over the heatsink in K. (a) Tier A. (b) Tier B. (c) Tier C. (d) Surface.

Fig. 15. 3-tier response mask for the MITLL Process calculated with a resistive mesh-based model for a heat load on a 2 \( \mu \text{m} \) wide transistor on Tier C. \( x \)- and \( y \)-axes are distance in \( \mu \text{m} \), \( z \)-axis is temperature rise over the heatsink in K. (a) Tier A. (b) Tier B. (c) Tier C. (d) Surface.

The runtime for a full resistive mesh-based simulation of the validation design was 11:26:39. The time for a Power Blurring HD simulation was on the order of 1 s. A more detailed breakdown of the runtimes is given in [30]. While the time to generate the response masks was significant, these calculations only need to be performed once. These responses can then be used for similar transistors throughout the design. While there is no benefit to using this approach on such a small design, these responses will be used to obtain a significant speedup at the full-chip scale in Section VI-C.

The results of the Power Blurring HD simulation of the SAR validation design are shown in Table II and Fig. 16. Comparing the full resistive mesh-based simulation to Power Blurring HD shows good agreement. The heat spreading effect of the backmetal shapes has a significant impact on the temperature profile of the surface of the chip; however, the temperature profile of the active layers for the circuit still shows significantly steep per-transistor changes. It is important to simulate the thermal profile of the circuit on the active layers as the individual transistors can be significantly hotter than the surface. This response remains hidden when only the top surface of the chip is examined.

B. Response Mask Variance

The exact shape of the response mask that is used to model a specific transistor will vary based on which transistor in the design is selected to create the response mask. Instead of computing individual response masks for every transistor in the circuit, the goal is to select an appropriate number of response masks to model the circuit such that the error in the final thermal profile is acceptable. The accuracy of Power Blurring HD is, therefore, dependent on the sampling of transistors to create response masks. The use of fill structures helps to equalize many areas of the chip and reduce the effect of the exact shape of local interconnect; however, it does not remove all variations. In general, circuits with lower variation
TABLE II
AVERAGE PER-TRANSISTOR CHANNEL TEMPERATURES FOR ALL TRANSISTORS IN THE SAR VALIDATION DESIGN AS CALCULATED BY A FULL RESISTIVE MESH-BASED MODEL AND BY A POWER BLURRING HD SIMULATION

<table>
<thead>
<tr>
<th>Tier</th>
<th>Average Channel Temp. Rise (K)</th>
<th>Individual Channel Temp. Rises</th>
<th>Max. Error (%)</th>
<th>Root Mean Square Error (K)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>17.54</td>
<td>16.44</td>
<td>12.9</td>
<td>1.33</td>
</tr>
<tr>
<td>B</td>
<td>24.13</td>
<td>23.46</td>
<td>3.95</td>
<td>0.721</td>
</tr>
<tr>
<td>C</td>
<td>35.03</td>
<td>33.92</td>
<td>7.94</td>
<td>1.53</td>
</tr>
<tr>
<td>Surface*</td>
<td>23.00</td>
<td>21.54</td>
<td>6.34</td>
<td>1.46</td>
</tr>
</tbody>
</table>

Individually channel temperatures were calculated by averaging the temperatures in all grid squares directly over that transistor’s channel on its active layer or on the top layer for the surface.

*Temperature rise over area bounded by Tier C channels.

Fig. 16. Comparison of simulation results for the SAR validation design between a full resistive mesh-based model simulation and a Power Blurring HD simulation. Responses are shown for the active layers of Tiers A, B, C and the surface of the chip. z-axis is temperature rise over the heatsink in K. (a) Tier A: resistive mesh. (b) Tier B: resistive mesh. (c) Tier C: resistive mesh. (d) Surface: resistive mesh. (e) Tier A: power blurring HD. (f) Tier B: power blurring HD. (g) Tier C: power blurring HD. (h) Surface: power blurring HD.

To test the variance in the SAR, response masks were generated for a sampling of transistors. Variation is expected between these response masks due to the variation in the size of active islands in particular around each transistor. Three 2 μm wide transistors were selected on each tier for a total of nine response masks. The selected transistors had a wide variety of active island sizes. Table III shows the sample mean and sample standard deviation.

The low variations seen in this paper indicate that the entire area of digital logic in the SAR (and hence the SAR validation sample design) can be appropriately modeled with single response masks per tier for each transistor size. This can also be generalized to mean that all designs using the same standard cell library as the SAR are a good fit for these response masks.

TABLE III
SAMPLE MEAN AND STANDARD DEVIATION OF THE AVERAGE CHANNEL TEMPERATURE ON TIER A, B, C AND THE SURFACE FOR HEAT LOADS PLACED ON TIER A, B AND C

<table>
<thead>
<tr>
<th>Heat Load on Tier</th>
<th>Statistic</th>
<th>Channel Temp. Rise on Tier</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Sample mean</td>
<td>0.099 0.994 0.588 0.337</td>
</tr>
<tr>
<td></td>
<td>Sample std. dev.</td>
<td>0.324 0.009 0.003 0.003</td>
</tr>
<tr>
<td>B</td>
<td>Sample mean</td>
<td>0.342 1.295 1.031 1.031</td>
</tr>
<tr>
<td></td>
<td>Sample std. dev.</td>
<td>0.020 0.633 0.032 0.025</td>
</tr>
<tr>
<td>C</td>
<td>Sample mean</td>
<td>0.342 1.129 12.118 3.203</td>
</tr>
<tr>
<td></td>
<td>Sample std. dev.</td>
<td>0.014 0.051 0.531 0.499</td>
</tr>
</tbody>
</table>
Fig. 17. Power Blurring HD simulation of the 3-D SAR. Power dissipation in the RAMs and ROMs is not considered. $x$ and $y$-axes are distance at μm, $z$-axis is temperature rise over the heatsink in K. (a) Surface. (b) Tier C. (c) Tier B. (d) Tier A.

Given a large enough sample of response masks for transistors in each region, two additional full-chip simulations could be done in order to establish the bounds of the per-transistor temperatures in the full-chip simulation: one using the response masks that yield the lowest temperatures and one using the response masks that yield the highest temperatures.

C. Full-Chip Power Blurring HD

The response masks generated for the SAR validation example were used for the full-chip SAR. The response masks use a grid size of 0.1 μm. The same grid size was used for the full-chip convolution.

The full-chip responses for Tiers A, B, C and the surface were calculated separately. The responses are shown in Fig. 17. With the use of 0.1 μm grid squares, sharp spikes (“tentpoles”) in the Tier B response are clearly visible. The spikes are caused by individual clock buffer cells with power dissipations that are significantly higher than those of the cells near them. The effect of these clock buffers can also be seen with smaller spikes in the corresponding locations on Tier C. The ability to properly capture the effects of these clock buffers requires a model that accurately describes the physical structure of the chip near these power sources at a resolution close to the junction scale [26]. Based on these simulations, it is clear that Power Blurring HD is able to work on the full-chip scale at a resolution fine enough to properly capture this effect.

Based on models for the CPU and the memory requirements of resistive mesh-based models, a full resistive mesh-based simulation of this design would have required 12 TB of memory for the matrix solve and 28 TB of memory for the extraction. The matrix solve alone would have required approximately 777 years of CPU time to complete. Instead, the full-chip simulation was performed with 23.8 GB of RAM in 4 h and 40 min of CPU time using Power Blurring HD. Power Blurring HD required 509 times less memory than a full-chip matrix solve on a resistive mesh-based model. The total amount of time needed to generate the response masks for the SAR validation example was approximately 36 h with a maximum memory usage of 13.2 GB. Even if the time to generate the response masks (which only needs to be done
VII. Asynchronous Multiplier/Divider Results

This section compares the Power Blurring HD approach to thermal measurements.

A. Chip Design

The asynchronous multiplier/divider chip [31] measures 3.76 mm × 3.76 mm and was fabricated in the three-tier MIT Lincoln Laboratory 3-D process. The total power consumption of the chip is approximately 1 W. The design consists of a 3 × 3 grid of evenly spaced power dissipating units on each tier for a total of 27 units. These units are seen in the die micrograph shown in Fig. 18 and appear as large, faint squares. Inside each of these units is a chain of frequency multiplier and divider cells where the majority of the power is dissipated. A total of 15 frequency multiplier and 10 frequency divider cells are placed back-to-back to create a chain. For each successive multiplier in the chain, the frequency of the signal propagating along the chain is doubled. Each divider in the chain then divides the frequency in half. The maximum operating frequency (1 GHz) is found in both the 15th multiplier cell and the first divider cell.

B. Measurement Setup

Infrared imaging passively measures the temperature profile based on the infrared radiation emitted from the chip and allows for nondestructive measurement of the chip. A Quantum Focus infrared camera was used for this measurement.
The chip was epoxied to a ceramic pin grid array package which was then placed into a socket on an FR4 circuit board which was attached to a large copper heatsink. In order to allow the best possible thermal path from the bottom of the chip to the heat sink, a pillar was added to the heatsink to connect directly to the bottom of the package. The use of a large heatsink beneath the circuit board allows the entire assembly to be placed on a heated chuck for measurement.

The test assembly for the asynchronous multiplier/divider chip was placed on a heated chuck directly underneath the infrared camera. The chuck was heated to 80 °C throughout the measurements. By raising the temperature of the chip, the amount of radiation generated by the chip becomes significantly higher than the background radiation level.

C. Power Blurring HD Model

One response mask was extracted for a single transistor on each tier for a total of three response masks. All of the response masks were generated using 1.8 μm wide transistors from the middle of the chain of frequency multiplier and divider cells. Each response mask was calculated to include 32 μm of the chip in both the x and y-directions from the transistor channel of interest. The response was calculated with a resistive mesh-based model using a lateral grid size of 0.1 μm.

D. Full-Chip Power Blurring HD

Full-chip analysis at a 0.1 μm grid size was performed using the Power Blurring HD in MATLAB. The simulation required 36.6 GB of memory and approximately 70 min of CPU time per profile on an 8-core Intel Xeon server with 32 GB of RAM. The wall time (i.e., the time that would be recorded by a clock on the wall) was just over 11 min per profile yielding a speedup of 6.30 times due to the parallelization of the 2-D convolution routine (conv2) in MATLAB.

Fig. 19 shows the temperature profile of the surface of the chip during operation as determined by both simulation and measurement. When the chip was powered off, the surface of the chip was measured to be a uniform 80 °C, i.e., the same temperature as the heated chuck. When the chip was powered on, a background temperature of 81.68 °C was measured. This temperature rise is likely due to the thermal resistance of the epoxy between the package and the chip, the package itself, the interface between the heatsink and the package, as well as the heatsink itself which was mounted without thermal paste on the heated chuck. Since the exact resistances of these interfaces are unknown, an offset of 1.68 °C was used to adjust the Power Blurring HD simulation so that it could be compared with the measurement. This approach is consistent with the 1-D thermal resistances that were used to model the packaging environment in [4] and [28]. For complex packages, an intrinsic error correction step, such as that used in [23], may also be useful.

The mean of the simulated profile over the four hottest cells in the upper middle unit was found to be approximately 21% higher than the mean for the same area in the measurement. This difference comes from discrepancies between the simulated and measured power dissipation profiles, the ideal and actual power sources, and the assumed thermal resistance and thermal conductivity of the materials used in the package. The heat from the power dissipation in the chip is transferred to the heatsink and the surrounding materials, and some of this heat is conducted to the ambient environment. The layout of the chip and the package, the placement of components on the chip and in the package, and the placement of the chip in the package and on the heatsink all affect the thermal behavior of the system. The simulation and measurement were performed at the same temperature, and the same substrate, but the actual temperature profile of the chip during operation may be different from the simulated profile due to the different materials and geometries used in the package.

Fig. 21. Power Blurring HD simulation results for the asynchronous multiplier/divider chip. The simulation has been adjusted to match the offset of the measurement and to match the average hotspot temperature so that the shape of the response can be compared. x and y-axes are distance in μm, z-axis is temperature rise over the heatsink in K. (a) Surface. (b) Tier C. (c) Tier B. (d) Tier A.
actual material properties and may also indicate that more response masks need to be used. A uniform multiplier of 0.825 was applied to the Power Blurring HD output in order to allow for easy comparison of the difference in shape between the simulated and measured temperature profiles.

Fig. 20 shows a comparison between the measured profile of the upper middle unit on Tier C and an adjusted simulation profile of the same area. The root mean square error for this region is 0.4926. The profiles show a good match indicating that Power Blurring HD has successfully predicted the shape of the response and the relative differences in temperature between the circuit elements, even though the actual values from the simulation underestimate the temperatures by 21% on average.

Thermal profiles were also generated for the full chip on the active layers of Tiers A, B, and C, as well as on the surface of the chip. Full-chip 3-D profiles are shown in Fig. 21. It is interesting to note that the maximum simulated channel temperature for the upper middle unit on Tier C is more than 40% higher than the maximum temperature for the corresponding area in the surface simulation.

Finally, it is important to note that while the asynchronous multiplier/divider chip (and in fact the SAR chip also) was arranged such that the areas modeled by the near responses never extended beyond the edge of the chip, this may not always be the case. In such a case, which is likely to arise in technologies that have significant lateral heat conduction paths or in designs where transistors are located close to the edges of the chip, the method of images examined in [19] could be used to account for the edge effects.

VIII. CONCLUSION

The degraded thermal path of 3DICs makes thermal analysis at the chip-scale an essential part of the design process. Performing an appropriate thermal analysis on such circuits requires a model with a junction-level fidelity; however, the computational burden imposed by such a model is tremendous. Two methods were presented for modeling 3DICs that are capable of maintaining this level of detail.

A resistive mesh-based thermal modeling approach was used to consider the full structure of 3DICs, including interconnect. It was found that the model converged when the grid size was at least as small as the minimum gate length. Using larger grid sizes led to underestimation of channel temperature. Furthermore, not modeling the exact orientation of transistor channels on active islands was found to yield incorrect temperature profiles in silicon-on-insulator processes.

Power Blurring HD was presented to extend the original Power Blurring approach in order to: 1) appropriately model structures found in SOI 3DICs, and 2) enable junction-level thermal simulation of SOI 3DICs at the full-chip scale with a runtime that is similar to common digital design tools. Power Blurring HD was estimated to yield three orders of magnitude of improvement in memory usage and up to six orders of magnitude of improvement in runtime for a 3 mm × 3 mm 3-tier circuit modeled with a 0.1 μm × 0.1 μm element size, as compared to directly solving the full-chip junction-scale thermal network. Measurement results were presented showing that Power Blurring HD is able to accurately determine the shape of the thermal profile of the surface of a 3DIC after adding a correction factor to adjust for a discrepancy in the absolute temperature values.

ACKNOWLEDGMENT

The authors would like to thank F. Akopyan, C. Otero, and R. Manohar for the asynchronous multiplier/divider chip, S. Dooley for the infrared camera measurements, A. Srinivasan and E. Cheng for modeling both chips in Gradient HeatWave-3DIC, and the MIT Lincoln Laboratory, Lexington, for providing access to their technology.

REFERENCES

169

Samson Melamed (M’11) received the B.S. degree in computer engineering from the University of Maryland, Baltimore County, Baltimore, in 2004, and the M.S. and Ph.D. degrees in electrical engineering from North Carolina State University, Raleigh, in 2007 and 2011, respectively. He is currently a Post-Doctoral Fellow with the National Institute of Advanced Industrial Science and Technology, Tsukuba, Japan. His current research interests include design and circuit design strategies for 3-D integrated circuits.

Thorlindur Thoroelsson (S’04–M’11) received the B.S. degree from Reykjavik, Iceland. He received the B.S., M.S., and Ph.D. degrees from North Carolina State University, Raleigh, in 2004, 2005, and 2011, respectively.

T. Robert Harris (S’04–M’11) received the M.S. and Ph.D. degrees in electrical engineering from North Carolina State University, Raleigh. He is currently with Dol Melins Technologies, Inc., Charlotte, NC. His current research interests include thermal modeling and management, multiphysics simulation, integrated circuit fabrication, and ultraviolet optics.

Shivram Priyadarshi received the B.Tech. degree in information and communication technology from the Dhirubhai Ambani Institute of Information and Communication Technology, Gandhinagar, India, in 2005, and the M.S.E.E. degree from North Carolina State University, Raleigh, in 2010, where he is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering. From 2005 to 2008, he was with Viager Logic, Fremont, CA, where he worked on physical IP development of embedded SRAMs and standard cells. His current research interests include electrothermal modeling and simulation of 3-D integrated circuits, parallel circuit simulation, and computer architecture.

Paul Franzon (SM’99–F’06) received the Ph.D. degree from the University of Adelaide, Adelaide, Australia. He is currently a Professor with the Department of Electrical and Computer Engineering, North Carolina State University, Raleigh. He is also with the Defense Science and Technology Organization, Adelaide, Australia, Telcos Australia, Adelaide, and two startups he cofounded. His current research interests include design and computer-aided design of integrated microsystems.

Michael B. Steer (S’76–M’78–SM’90–F’99) is currently the Lampre Distinguished Professor of electrical engineering with the Department of Electrical and Computer Engineering, North Carolina State University, Raleigh. He has authored 450 papers and written three books.

W. Rhiht Davis (S’92–M’02–SM’10) received the B.S.E.E. degree from North Carolina State University, Raleigh, in 1994, and the M.S.E.E. and Ph.D.E.E. degrees from the University of California, Berkeley, in 1997 and 2002, respectively. He is currently an Associate Professor with the Department of Electrical and Computer Engineering, North Carolina State University. His current research interests include developing methodologies, computer-aided design tools, and circuits for systems-on-chip in emerging technologies.