Junction-level Thermal Extraction and Simulation of 3DICs

Samson Melamed, Thorlindur Thorolfsson, Adi Srinivasan[†], Edmund Cheng[†], Paul Franzon and Rhett Davis
Dept. of Electrical and Computer Engineering, North Carolina State University, Box 7911, Raleigh, NC 27695
†Gradient Design Automation, 4633 Old Ironsides Drive, Suite 200, Santa Clara, CA 95054
{slmelame,trthorol,paulf,wdavis}@ncsu.edu
{adi,edcheng}@gradient-da.com

***Abstract*—In 3DICs heat dissipating devices are stacked directly on top of each other leading to a higher heat density than in a comparable 2D chip. 3D integration also moves the majority of active devices further away from the heatsink. This results in a degraded thermal path which makes it more challenging to remove heat from the active devices. Gradient FireBolt was used to perform an appropriate 3D thermal analysis on a 1024-point, memory-on-logic 3DIC FFT processor for synthetic aperture radar (SAR). The chip was simulated with a spatial resolution of 80 nm, and was modeled to include the effect of each line of interconnect, as well as each via and fill structure exactly as drawn in the layout. Large isolated temperature spikes were found near groups of clock buffers at the edge of the SRAMs on the middle tier. It was found that lowering the simulation resolution and using composite thermal conductivities failed to accurately predict the location of these tentpoles.**

I. Introduction

The scaling of silicon technology has steadily increased power density such that a cluster of high performance devices can produce small, high temperature hotspots in a 2D chip [1]. In 3DICs heat dissipating devices are vertically stacked, leading to larger spatial variations in power-density, and hence larger thermal gradients, than in comparable 2D chips. High thermal gradients have been tied to unexpected electrical failures, undetected by traditional methods where all transistors are set to the maximum temperature [2].

3D integration moves active devices (excluding those on the lowest tier) further away from the heatsink. This results in a degraded thermal path which makes it more challenging to remove heat from the active devices. In many cases the thermal issues that arise because of this must be taken into account during the design cycle in order to realize a working chip. Several research efforts have taken thermal effects into account, most often in the form of standard cell placement [3], [4]. Accurate modeling of the thermal profile of the chip has become a critical step in allowing designers to consciously consider the effect of heat dissipation and power density.

Existing thermal-simulation methods, when applied to a full-chip, reduce the computational complexity of the problem by homogenizing the materials within a layer, limiting the extent of an eigenfunction expansion, or ignoring sources' proximity to boundaries. These simplifications render their results inaccurate at fine length-scales, on wires, vias, or individual transistors.

In this paper we present a full-chip thermal simulator, Gradient FireBolt [5], a technique for generating the necessary input from layout, and the results obtained from analyzing a 3D FFT processor with this simulator.

The remainder of this paper is organized as follows. Sec. II provides an overview of thermal modeling issues, provides background on thermal simulation techniques, and describes the FireBolt [5] simulator. Sec. III describes the sample design and the technology that it was designed in. Sec. IV describes the flow for extracting accurate power source information from the sample design. Sec. V shows the results from low, medium and high resolution simulations of the sample design. Sec. VI provides tips for designers to decrease the likelihood of encountering thermal problems.

II. Simulation Techniques

A. Overview

The effects of temperature on semiconductor devices are well documented [6], including reduced carrier mobility, MOS threshold $|V_T|$ reduction and increased sub-threshold leakage, among others. Electromigration failure rates in metal interconnects and vias increase with temperature [7], [8].

Even in a single die, the characterization and analysis of such effects requires accurate temperatures of transistors, metal segments and vias under given operating conditions. When die are thinned and stacked, as the tiers of a 3DIC, the new vertical proximity of power sources, and general degradation of thermal paths only increases the need for high-resolution thermal analysis.

Accurate computation of temperature at the length-scales of devices and interconnects requires a detailed accounting of the heat flow from the power-sources through the nanometer-scale layout within the chip. In 3DICs, this includes the thermal effects of through-silicon vias (TSVs) which can cause significant local temperature variations. As will be shown, ignoring layout details by use of a composite or homogenized model, or by use of a coarse mesh, can result in significant errors in temperature.

B. Background

Existing intra-die temperature modeling and computation techniques are varied, a representative sample being [2], [9]–[12]. Discretization or mesh based methods [2], [12] can model

978-1-4244-4512-7/09/$25.00 ©2009 IEEE

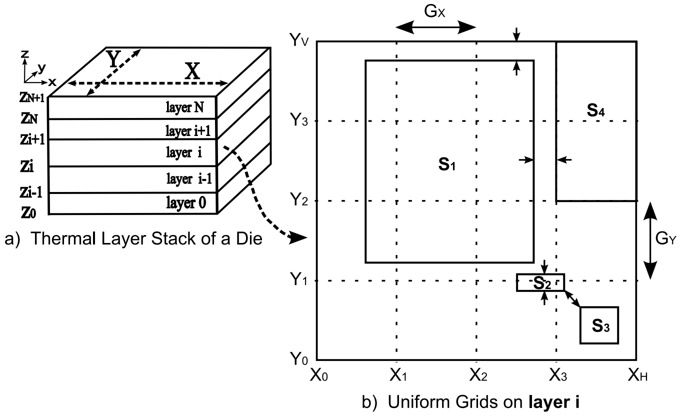


Fig. 1. Uniform or coarse grids on a layer

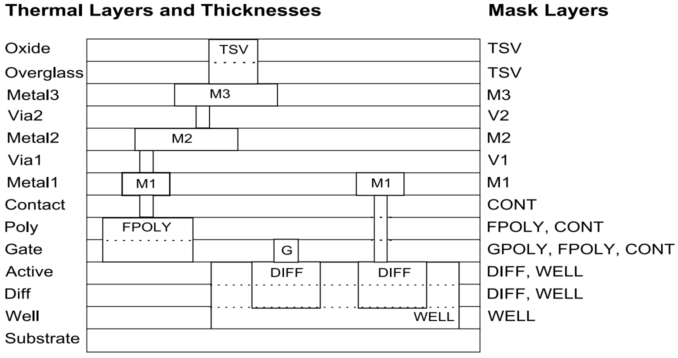


Fig. 2. Thermal-layer stack using mask-layer data

the heterogeneous material distribution implied by the chip's layout-data. This may be prohibitively complex on a full-chip scale [2]. Green's function based methods [9]–[11] are faster, with some simplifying assumptions.

Most methods (e.g. [2], [9]–[11]) use a Fourier heat transport model, as expressed by the heat diffusion equation [13], which in steady state, in Cartesian co-ordinates is

$$\nabla \cdot [k(\mathbf{r}, T)\nabla T(\mathbf{r})] + P_V(\mathbf{r}) = 0 \tag{1}$$

where $\mathbf{r} \equiv (x, y, z)$ in (m), T is the temperature (K), k is the thermal conductivity (W/(mK)), and P_V is the power density (W/m^3). Dirichlet, Neumann or Robin boundary conditions (BCs) [13] may be applied at the six die faces. P_V is considered invariant with temperature, assuming power values may be iteratively updated in an electro-thermal loop, with an electrical simulator computing powers at the updated temperatures. Sub continuum heat-transport is modeled using the Boltzmann Transport Equation (BTE) only for intra-device heating (the inter-device effect is negligible) via a hybrid Fourier-BTE model in [12].

Methods [2], [9]–[12] make the linearizing assumption

$$\partial k(\mathbf{r}, T)/\partial T \equiv 0 \tag{2}$$

The die may be modeled as a stack of layers as in Fig. 1(a). An arbitrary layer i extends vertically from Z_i to Z_{i+1}. In general the layer material may be inhomogeneous and anisotropic, with the material composition determined by the layer's layout geometries.

Green's function methods [9]–[11] further assume layer homogeneity and isotropy, so

$$k(\mathbf{r}) \equiv k_i \tag{3}$$

in layer i.

Then Eq. 1 reduces to Poisson's equation in any thermal layer i

$$k_i \nabla^2 T(\mathbf{r}) = -P_V(\mathbf{r}) \ \text{ for } \ z \in [z_i, z_{i+1}] \tag{4}$$

[9]–[11] express the temperature as a Fourier expansion on an implied uniform grid. As in Fig. 1(b), if the uniform grid has H horizontal intervals, and V vertical intervals, the Fourier expansion for one of the HV regions' temperature (see [9], [11]) contains HV terms, with the finest resolution determined by the grid spacings $G_X = X/H$ and $G_Y = Y/V$ given a die size of X by Y. While this may sufficiently sample the larger power-sources (S_1, S_4), smaller sources (S_2, S_3) are not well sampled. In Fig. 1(b), arrows indicate where temperature variations might be missed: on smaller power-sources (S_2), small gaps between power-sources (S_1–S_4 and S_2–S_3) and small gaps between power-sources and boundaries (S_1 to top-edge). The grid spacings (G_X, G_Y) must be on the order of these distances for a converged solution. In the SAR design, a resolution around 100 nm in the 3 mm chip is required, so $H = V \cong 30000$, implying 9×10^8 terms in the Fourier expansion, which is intractable for [9], [11].

Similar complexities arise in mesh-based solvers using a relatively coarse mesh, due to the number of power-sources and layout shapes which must be sufficiently meshed (i.e. sampled) to ensure a converged solution.

Further inaccuracies result from Green's function based methods' (see [9]–[11]) assumptions of material homogeneity within a layer (or the whole die), which ignore the significant variations in heat transport due to variations in the layout.

[12] uses a hierarchical adaptive grid to sample the power-sources. However none of the previous approaches samples the layout at its length scales, which is needed to accurately compute the heat transport from the power-sources through the layout including the TSVs and metal fill.

C. Gradient FireBolt

The FireBolt [5] thermal solver computes the steady state temperature from Eq. 1 modeling the detailed layout, with temperature dependent conductivity (avoiding the approximations in Eqs. 2 and 3). Initial meshless computations are used to efficiently discretize the 3DIC model, within a multi-level hierarchical solver. Sub continuum heat transport is modeled as needed.

FireBolt reads a layout database in GDSII or OpenAccess. A technology file defines material conductivities, and a thermal layer stack. Fig. 2 shows a cross section of the bottom die

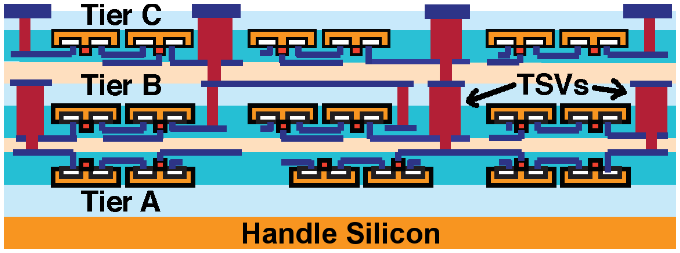


Fig. 3. The MIT Lincoln Laboratory's 3D process stackup [14]. The handle silicon for the bottom tier remains intact. The handles from the middle and upper tiers are thinned, and then wafer-bonded upside down. Through-silicon vias (TSVs) are used to connect signals between the tiers. Note: Layer thicknesses are not to scale.

in a 3DIC stack, in a process similar to that of the SAR design. In Fig. 2, a stack of "thermal layers" is defined, each with a name and a thickness. The materials in a thermal layer are determined by layout data on mask-layers, each of which is also associated with a material (not shown in Fig. 2). If the M2 mask is associated with Aluminum, the Metal2 layer will be made of Aluminum wherever a shape exists on the M2 mask, and will be made of oxide where no layout-shapes exist. Multiple masks may be associated with a thermal-layer (e.g. thermal-layer Gate is built from layout on mask-layers GPOLY, FPOLY and CONT). Power sources are positioned in the thermal-layers, e.g. transistor power-sources would be positioned in the Active thermal-layer.

This representation enables the rapid preparation of design and process data for thermal analysis.

We used FireBolt to compute the temperature profile of the SAR 3DIC design, at different spatial resolutions. The finest spatial resolution used in the converged solution was 80 nm and temperatures were resolved to $\leq 0.1°$C. For comparison, non-converged low-resolution and moderate-resolution temperature profiles were also computed.

III. Sample Design and Technology

The MIT Lincoln Laboratory (MITLL) process uses three wafer-integrated tiers. A diagram of the process stackup is shown in Fig. 3. Each tier is fabricated with a 675 μm thick silicon handle beneath the buried oxide. For the middle and upper tiers, the handle is thinned away completely, and the tiers are then flipped upside down and bonded to the tiers beneath it. After integration, there is a single silicon handle at the bottom, followed by three tiers of logic at the top. No handle silicon remains between the logic devices, and the only available heat conduction paths from the upper tier to the lower tier is through interlayer dielectric (SiO_2) and 3D vias. 3D vias in the process can be freely placed, and connect between metal layers in the tiers.

The floorplan of the SAR is shown in Fig. 5. The floorplans from left to right are for the bottom, middle and upper tiers (Tiers A, B and C.) Both tiers A and C contain four processing elements, along with a set of ROMs in the middle. The middle tier contains 32 SRAMs, and controller logic. This arrangement allows the memories to be easily accessible from processors on both tiers. Fig. 4 shows the location of 3D vias. Due to the architecture of the SAR, the tiers are densely connected with a relatively uniform layout of 3D vias.

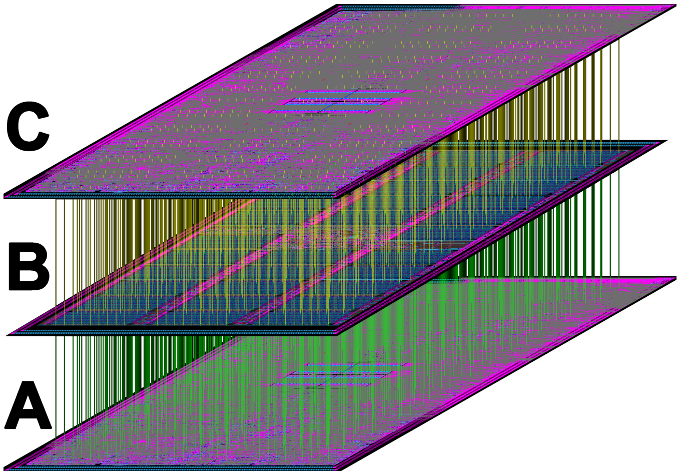


Fig. 4. Stackup of the 3D SAR, showing connectivity between the tiers. 3D vias are shown in yellow (connecting between tiers C and B) and green (connecting between tiers B and A.) Note: 3D via dimensions are not shown to scale.

The SAR contains a total of 786,147 power sources. The dimensions of layout objects range from 0.2 μm to 690 μm. The total power dissipated in the SAR is 704 mW. The size of each die (or tier) is 3 mm in the $x-$ direction and 3 mm in the $y-$ direction. The variation of power with temperature is not modeled.

IV. Model Extraction

In this section we describe the methodology for extracting per-transistor power values and locations from the design.

Obtaining accurate thermal simulation results relies not only on being able to accurately model the physical structure of the circuit, but also on the ability to accurately extract power dissipation information. This includes describing both the location of power dissipation as well as the amount of power dissipated at each location.

In order to take advantage of a high resolution thermal simulator, power values must be extracted in a way that does not distort their location. This is particularly important in processes where the primary heat conduction paths are through the metal layers. If power sources are distorted (e.g. specifying a single power dissipation area for a floorplan block), it will appear as if power is being directly dissipated in both the metalization and the dielectric. This will force heat to flow through the dielectric, which may provide erroneous results. Using single average-power numbers over large blocks also distorts the concentration of heat generation, leading to significant miscalculations in areas where the temperature changes quickly.

The extraction flow (shown in Fig. 6) requires an OpenAccess layout database, synthesized verilog and parasitic information for the chip. The verilog is simulated in ModelSim in order to generate a switching activity information file

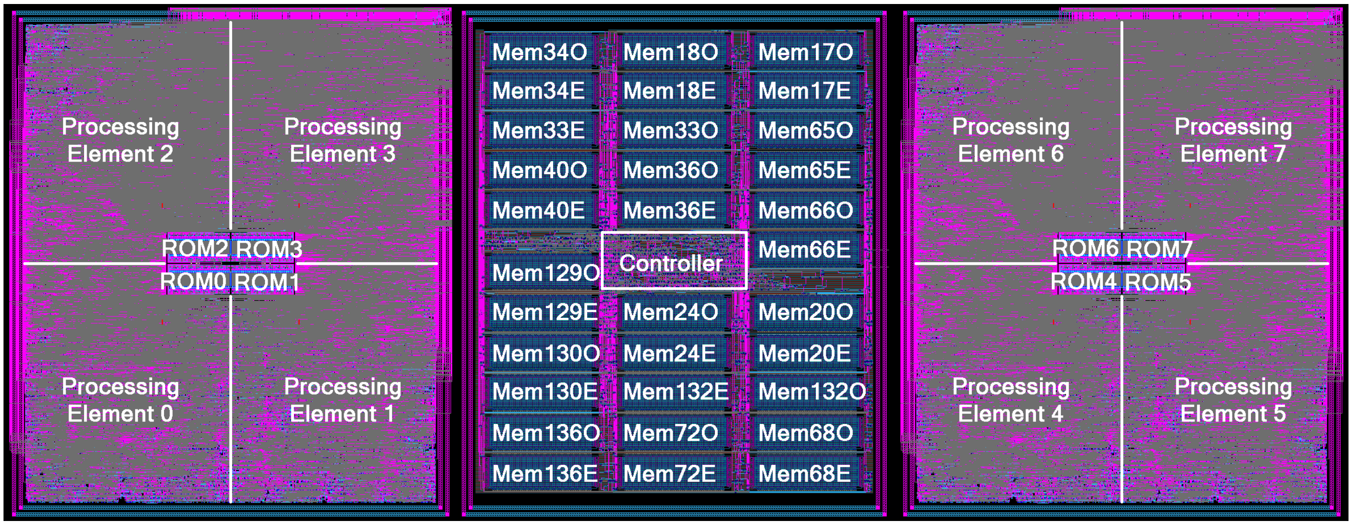


Fig. 5. Floorplan of the 3D SAR. Shown from left to right: Tier A (bottom), Tier B (middle), Tier C (top).

(SAIF). The SAIF file is then read into Synopsys Design Compiler along with the parasitic information file. Per-tier parasitic information is generated by Encounter at the end of the place and route phase, and saved as a SPEF. Synopsys Design Compiler uses the verilog, SAIF and SPEF along with the standard cell library's characterization information to determine per-cell average power dissipation values.

The Synopsys power report is then fed into a Perl script, `statdynpower.pl`, to format the data for PowerNote. PowerNote takes two separate power reports per tier – one for static power dissipation, and one for dynamic power dissipation – and back-annotates per-cell power values onto each instance in the OA layout database. These values are stored as properties on each standard cell instance.

In order to enable the extraction of per-transistor power values and locations, the standard cell library is processed with a custom Skill script (`findxistor.il`). The script creates shapes on active for the channel of each transistor. These shapes are then annotated with an ACTIVEDEVICE property to indicate that they are a power dissipation region, and a DEVICEWEIGHT property, to indicate the percentage of the total cell power that is dissipated in each region. For simplicity, the power is assumed to be evenly distributed across all transistors in the cell. An additional property, DEVICECOUNT, is then added onto the cell that indicates the total number of power dissipating areas.

The power values for the SRAM and ROM cells were annotated by hand, since they were not implemented with standard cells. By examining the design, it was found that each memory cell is accessed once per cycle, alternating between a read and a write operation. Spice simulations were performed for the memories, and an average power value was extracted assuming a read operation followed by a write operation. Average values for each SRAM and ROM block were calculated and back-annotated.

A custom Python script (`devicepower.py`) is then used to extract the power dissipation values and locations. This script iterates through all instances in the design, and uses the weights on the cell's power dissipation shapes along with the power dissipation values marked on the instance to determine the power dissipation for each transistor. These values, along with the outlines of the channels, are then written out to files for use by FireBolt.

V. Results

Three simulations of the SAR were performed with Gradient FireBolt. The first simulation ("low resolution") had a maximum element size of 94 μm and an adaptive minimum element size of 23.5 μm. This simulation used a single composite thermal conductivity for each layer in the stackup. The exception was 3D vias, which were modeled at full fidelity. Power sources were merged to be on the same scale as the model elements. The second simulation ("medium resolution") had a maximum element size of 23.5 μm and an adaptive minimum element resolution of 2.9 μm. This simulation used the full-chip layout, including all physical structures such as wires, vias and fill patterns. The third simulation ("high resolution") was allowed to converge with a maximum element size of 7 μm, a minimum element size of 80 nm and a maximum temperature change of 0.1°C. For the last simulation, the spatial and temperature resolutions were dynamically adjusted throughout the chip, to ensure that all areas of the chip simultaneously met or exceeded both resolutions. This simulation also used the full-chip layout.

Table I shows a summary of the simulation results. Plots of the temperature profiles on the active regions for all tiers are shown in Fig. 7. The simulation assumes that the bottom of the chip's handle silicon is attached to an ideal heatsink. This allows a prescribed temperature to be used as a boundary condition. A prescribed temperature of 27°C was selected.

The thermal profile of the active layer of the bottom tier is shown in Figs. 7(a), 7(d) and 7(g) for the low, medium and

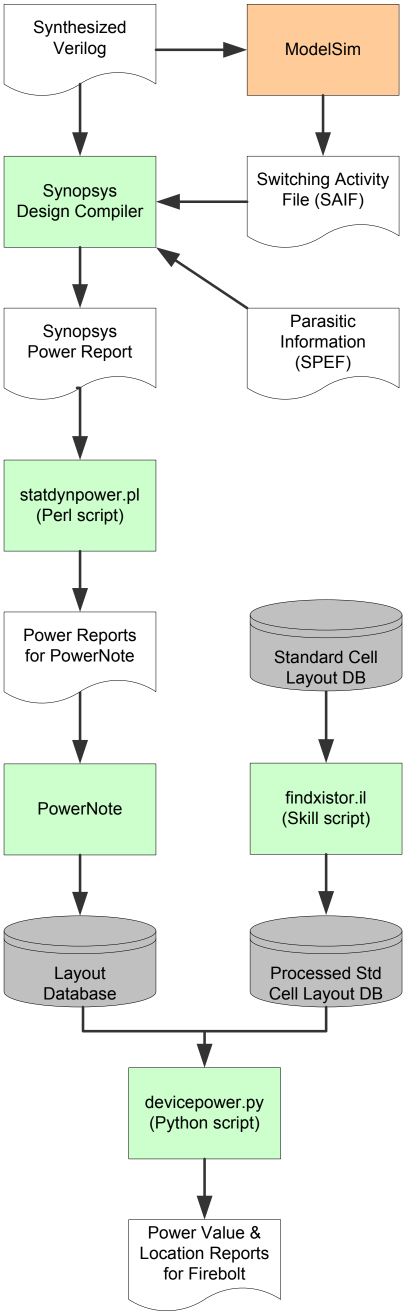

Synthesized Verilog
ModelSim
Synopsys Design Compiler
Switching Activity File (SAIF)
Synopsys Power Report
Parasitic Information (SPEF)
statdynpower.pl (Perl script)
Power Reports for PowerNote
Standard Cell Layout DB
PowerNote
findxistor.il (Skill script)
Layout Database
Processed Std Cell Layout DB
devicepower.py (Python script)
Power Value & Location Reports for Firebolt

Fig. 6. Flow diagram for the extraction of per-transistor power values and locations. This flow requires Synthesized Verilog as an input and produces the report files required to run a simulation with Gradient FireBolt.

TABLE I
COMPARISON OF SIMULATION RESULTS

	Low	Medium	High
Minimum Element Size (μm)	23.5	2.9	0.08
Maximum Element Size (μm)	94	23.5	7
Runtime (approx.)	6 min	150 min	90 hours
Memory Usage (approx.)	2.5G	8.5G	<64G
Maximum Temperature Rise (°C)	1.4998	24.8286	24.7341
Temperature Rise at u72_0 (°C)	1.1661	17.1954	24.7341

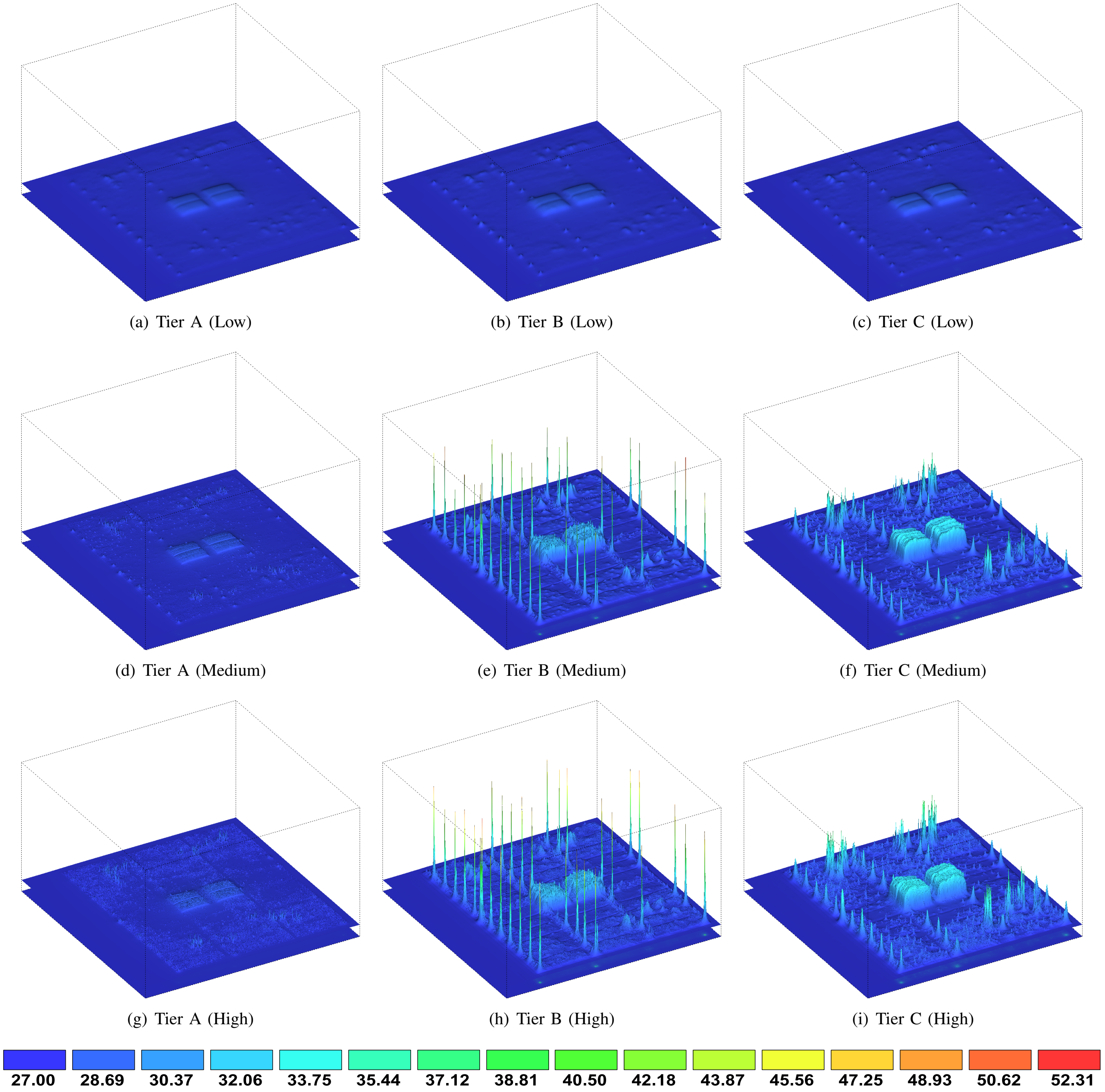
high resolution simulations, respectively. The bottom tier is unique in the fact that its handle silicon has not been thinned. In addition to being directly connected to the heatsink, the handle silicon has a high thermal conductivity (as compared to the interlayer dielectric), and serves as a heat spreader. This results in the bottom tier having a significantly better thermal conduction path than the upper tiers. In general, little difference is seen between the various resolution simulations for this tier.

Figs. 7(b), 7(e) and 7(h) give the thermal profile of the active layer of the middle tier for the low, medium and high resolution simulations, respectively. Fig. 8 shows a closeup of the high resolution simulation of tier B. Significant differences between the various resolution simulations are clearly seen. Due to the layout of this tier, clusters of clock buffers are found at regular intervals along the edges of the memories. The high and medium resolution simulations show dramatic tentpoles at these locations. The low resolution only shows bumps in the profile for these areas.

Figs. 7(c), 7(f) and 7(i) give the thermal profile of the active layer of the top tier for the low, medium and high resolution simulations, respectively. For this tier, the low resolution simulation once again fails to capture the tentpoles that are clearly evident in the medium and high resolution simulations.

Fig. 9 shows a histogram of the transistor temperatures obtained by all three simulations. While the low resolution simulation clearly underestimates the temperature profile, the medium resolution simulation matches well for temperatures < 35°C. Fig. 10 shows the ratio of the temperature calculated with the medium resolution simulation to that of the high resolution simulation. This figure only includes points where the high resolution simulation reported a temperature over 35°C. Numbers less than one indicate the medium resolution simulation is underestimating the temperature, while numbers over one indicate that it is overestimating the temperature. A total of 684 transistors fell into this category, with 101 (14.8%) being overestimated, and 583 (85.2%) being underestimated. The worst cases were 50.1% and 130% of the high resolution temperatures. The medium resolution simulation was able to come within 2× the value of the high resolution simulation for all transistors, while allowing a 36× decrease in runtime and a 7.5× decrease in memory usage.

Based on an analysis of the high resolution profiles, clusters of clock buffers were found to be the major source of tentpoles in this design. Accurately modeling these tentpoles requires detailed information about both the location of power dissipating devices, as well as the amount of power dissipated. The proposed standard-cell extraction flow is able to capture both the exact location of the transistors and the per-cell average power dissipation values. Since the activity in clock buffers remains constant and is spread evenly among the transistors, a single average power value provides a good approximation for these cells.



(a) Tier A (Low)

(b) Tier B (Low)

(c) Tier C (Low)

(d) Tier A (Medium)

(e) Tier B (Medium)

(f) Tier C (Medium)

(g) Tier A (High)

(h) Tier B (High)

(i) Tier C (High)

(j) Temperature Scale

Fig. 7. Low, medium and high resolution temperature profiles for the 3D SAR, computed by Gradient FireBolt. The results show the profile at the active layers in: the bottom tier (Tier A), the middle tier (Tier B), and the uppermost tier (Tier C). The bottom of the chip is attached to an ideal heatsink, with a prescribed temperature of 27°C.

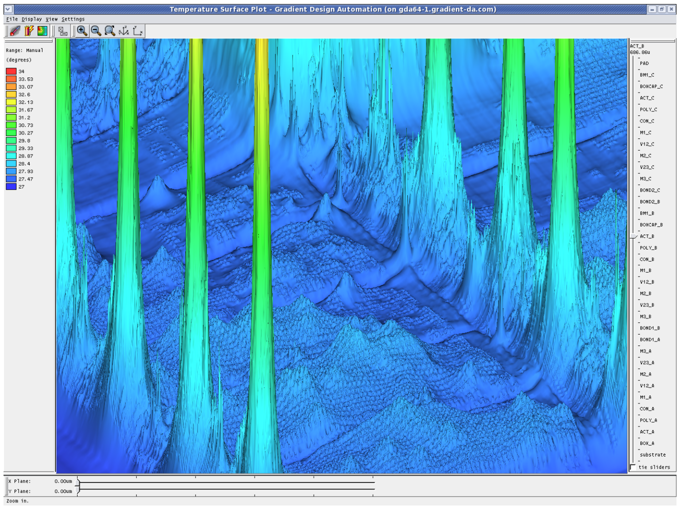


Fig. 8. A closeup of the analysis of the 3D SAR with Gradient FireBolt, showing the level of detail that is provided by a high resolution simulation. The region shown is on the middle tier (tier B). The tentpoles caused by clock buffers and the effects of wires are clearly seen at this resolution.

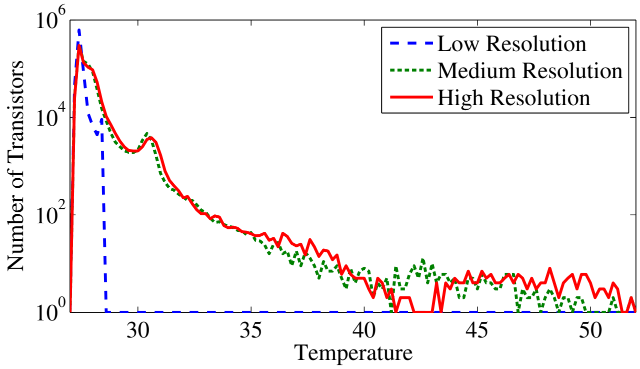


Fig. 9. Histogram of the SAR's transistor temperatures as determined by the low, medium and high resolution simulations.

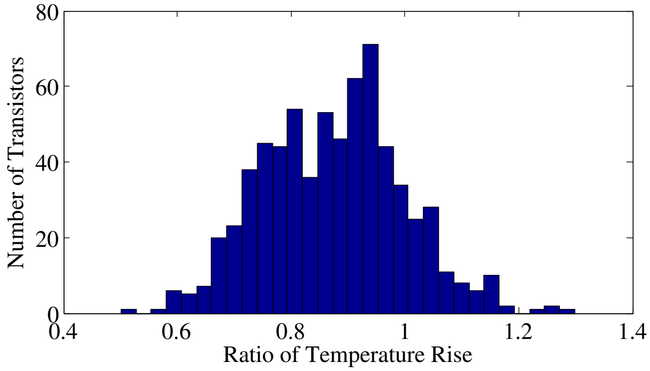


Fig. 10. Histogram of the ratio of transistor temperature rise as determined by the medium resolution simulation vs. the high resolution simulation, for transistors that were above 35°C in the high resolution profile. Numbers less than one indicate that the medium resolution simulation underestimated the temperature rise.

VI. Conclusion

Medium and high resolution thermal analysis, when coupled with accurate power values and layout information, is able to capture tentpoles in the thermal profile. Low resolution simulations that use composite thermal properties obscure the tentpoles such that it is difficult to determine which areas will cause tentpoles.

The bottom tier was found to be highly insensitive to thermal issues due to its close proximity to the heatsink. If feasible, power hungry devices should be placed on this tier, to lower the likelihood of having thermal issues.

The temperatures determined by 3D thermal placement tools need to be critically analyzed. As seen in the low resolution results, simulations that do not take into account the full structure of the chip and the exact placement of heat conduction paths are not sufficient to locate trouble spots for these circuits. It should be noted that these tools may prove more accurate for 3D bulk processes, where the additional silicon handles provide an improved path for heat spreading.

Acknowledgement

This work was funded by DARPA under agreement number FA8650-04-C-7127, managed by the Air Force Research Lab. Additional funding was provided by the National Science Foundation under grant number 643700.

References

[1] D. J. Frank, "Power-constrained CMOS scaling limits", IBM J. Res. & Dev., Vol. 46, No. 2/3, pp. 235-244, Mar/May 2002

[2] Y. Cheng, P. Raha, C.C. Teng, E. Rosenbaum and S.M. Kang, "ILLIADS-T: An Electrothermal Timing Simulator for Temperature Sensitive Reliability Diagnosis of CMOS VLSI Chips", Comp. Aided Design of Circuits and Systems, Vol 17., No 8, Aug 1998

[3] J. Cong, J. Wei and Y. Zhang, "A thermal-driven floorplanning algorithm for 3D ICs", IEEE/ACM International Conference on Computer Aided Design, pp 306–313, 2004.

[4] B. Goplen and S. Sapatnekar, "Efficient thermal placement of standard cells in 3D ICs using a force directed approach", International Conference on Computer Aided Design, pp 86–89, 2003.

[5] FireBolt (Nanoscale Full-Chip Thermal Simulator), Gradient Design Automation, Inc. http://www.gradient-da.com

[6] S. M. Sze and K. K. Ng, Physics of Semiconductor Devices, 3rd ed., 2006

[7] J. R. Black, "Electromigration Failure Modes in Aluminium Metallization for Semiconductor Devices", Proc. IEEE, Vol. 57, No. 9, pp. 1587-1594, Sep. 1969.

[8] J. R. Black, "Electromigration – A brief survey and some recent results", IEEE T. Electron Devices, Vol 16, Issue 4, pp 338 - 347, April 1969

[9] Y. Zhan and S. S. Sapatnekar "Fast Computation of the Temperature Distribution in VLSI chips using the Discrete Cosine Transform and Table Lookup", ASPDAC, Jan 2005

[10] B. Wang and P. Mazumder, "Accelerated Chip-Level Thermal Analysis Using Multilayer Green's Function", IEEE T. Comp. Aided Design of Circuits and Systems, Vol 26., No 2, pp325-344 Feb 2007

[11] D. Oh, C. C. P. Chen and Y. H. Hu, "3DFFT: Thermal Analysis of Non-Homogeneous IC using 3D FFT Green Function Method", 8th Int. Symp. Quality Elect. Design, 2007

[12] N. Allec, Z. Hassan, L. Shang, R. P. Dick, R. Yang "ThermalScope: Multi-scale thermal analysis for nanometer-scale integrated circuits", pp 603-610, ICCAD 2008

[13] E. R. Eckert and D. Eckert, "Analysis Of Heat And Mass Transfer", eq (1-14), p.11, CRC Press, 1986

[14] MIT Lincoln Laborary Low-Power FDSOI CMOS Process Design Guide, September 2008.