Design and CAD for 3D Integrated Circuits

Paul D. Franzon, W. Rhett Davis, Michael B. Steer, Steve Lipa, Eun Chu Oh, Thor Thorolfsson, Samson Melamed, Sonali Luniya
NC State University,
ECE, Box 7911
Raleigh, NC 27695
+1.919.515.7351
paulf@ncsu.edu

Tad Doxsee, Stephen Berkeley, Ben Shani, Kurt Obermiller
PTC Inc
140 Kendrick St.
Needham, MA
+1.781.379.5706
kober@ptc.com

ABSTRACT
High density Through Silicon Vias (TSV) can be used to build 3DICs that enable unique applications in computing, signal processing and memory intensive systems. This paper presents several case studies that are uniquely enhanced through 3D implementation, including a 3D CAM, an FFT processor, and a SAR processor. The CAD flow used to implement for these designs is described. 3DIC requires higher fidelity thermal modeling than 2DIC design. The rationale for this requirement is established and a possible solution is presented.

Categories: B.7 Integrated Circuits; B.7.1 Types and Design Styles; VLSI

General Terms
Design, Verification.

Keywords
3DIC, Through Silicon Via, TSV, Thermal Modeling

1. INTRODUCTION
There is a growing consensus that there are several mainstream circumstances which justify 3D integration with high-density Through-Silicon Vias (TSV) [1]. A list of these of potential drivers for 3D integration is provided in Table 1. Many of these drivers involve complex integration of logic, memory and logic, or RF and logic. Several of these applications, especially miniaturization and mixed technology integration, have been well explored and do not often require high density TSVs; low-density vias suffice. In this paper, we explore applications that are enabled by high density TSVs, and are enhanced by the interconnect delay, memory bandwidth and power improvements they provide.

After exploring applications, we outline the CAD flow used in the synthesized digital problems explored. Finally, we discuss the potential issue of thermal design. Each additional layer of silicon devices proportionally increases the heat density and the attendant cooling problem. This means that more accurate thermal modeling is required for 3DIC design. In particular, more details about local thermal conductivity, including anisotropic details must be included in the model in order to get useful results. This is the final topic explored in this paper.

Table 1. Potential drivers for 3D integration.

Driving Issue	Case for 3D	Caveats
Miniaturization	Stacked memories. "Smart dust" sensors.	For many cases, stacking and wire-bonding is sufficient
Interconnect Delay	When delay in critical paths can be substantially reduced through 3D integration.	Not all applications will have a substantial advantage
Memory Bandwidth	Logic on memory can dramatically improve memory bandwidth	While memory bandwidth can be improved dramatically, memory size can only be improved linearly
Power Consumption	In certain cases, a 3D architecture might have substantially lower power over a 2D	Limited domain. In many cases, it does not.
Mixed Technology (Heterogeneous) Integration	Tightly integrated mixed technology (e.g. GaAs on silicon, or analog on digital) can reduce common mode noise, or provide for integrated electronics for non-CMOS imagers.	Though might justify 3D integration, not all examples might justify through-wafer vias.

DAC 2008, June 8–13, 2008, Anaheim, California, USA

2. APPLICATION EXPLORATION

When is it advantageous to go vertical and when is it not? Stacking two wafers together and integrating them with vertical vias is not cheap. As a rough rule of thumb, the additional processing cost is about equivalent to that of adding two additional layers of metal interconnect. This cost is even higher if individual dice are stacked. This cost must be justified through performance gains or cost savings elsewhere in the system. This section summarizes results obtained in explorations performed by our group.

We explored a number of interconnect-limited applications. An interconnect bound circuit is the Ternary Content Adressable Memory (TCAM). Remapping a TCAM onto 3D can provide a 23% power improvement because of the reduction in match line capacitance that can be achieved (Table 2) [2].

Table 2. Comarison of TCAM Array Energy-per-comparison

	Single-Tier	3 Tiers	Improvement
Match-Line	2.9 pJ	2.1 pJ	28%
Total	8.0 pJ	6.2 pJ	23%

Results obtained using two other practical examples explored at NCSU are summarized in Figure 1 [3]. This figure compares data taken from two designs executed in the Lincoln Labs 3-tier process. One is a Fast Fourier Transform (FFT) [3]. The other is a dual core processor – an Open Risc Processor System On a Chip (ORPSOC) [4]. In this study, the performance benefits of 3D integration were compared with those of technology scaling. In these examples, 3D integration provided about the same performance advantage of two generations of technology scaling – a very compelling case.

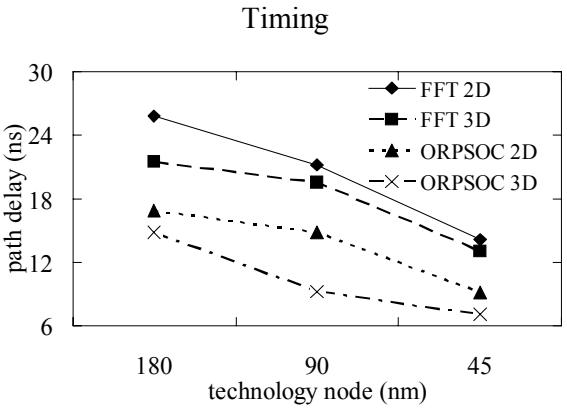


Figure 1. Improvement in path delay achieved using a 3D 180 nm technology, vs. technology scaling alone. 3 metal layers were assumed for each silicon layer. FFT=Fast Fourier Transform. ORPSOC = Open core RisC Processor System on a Chip. Adding two additional silicon layers is roughly equivalent to two generations of technology scaling.

For many end applications, the demand for memory bandwidth is growing rapidly. In many cases, this is due to the increased use of multi-core processors. With the addition of each processor comes a similar requirement for increasing memory bandwidth. It is predicted that by 2010 a 32-core CPU will require 1 TBps of off-chip memory bandwidth [5]. Other Applications that are likely to benefit form logic-on-memory include digital signal processing, graphics, and networking. For example, the demand for memory bandwidth for a Synthetic Aperture Radar (SAR) processor, changes as a function of desired image resolution. It would be very difficult to supply the memory bandwidth required for a high resolution 3D SAR with conventional memories. This is one application where a logic-on-memory structure could be particularly advantageous. Such stacking could permit both an increase in memory bandwidth, *and* a decrease in memory power consumption. The bandwidth increase comes about through the appropriate use of Through-Silicon Vias. The potential for power reduction arises through the ability to reduce the energy overhead of each memory access. Modern DRAMs are optimized to fill cache lines in processors. In a DMA application, such as many in DSP, different optimizations are possible, and the potential for power reduction is significant. Reducing memory power does increase the area of the memory, but one advantage of 3DIC implementation is that the total interconnected silicon area can be increased without a loss of bandwidth.

We investigated the advantages of implementing a SAR processor capable of producing three dimensional images using 3DIC integration technologies. A SystemC model of the processor was built and tested. The bandwidth and memory requirements of the processor are summarized in Table 3. Three alternative implementations were then explored and compared: the first using off-the-shelf components available today; the second assuming an energy-optimized 2DIC design built in a 45 nm process, and the third an energy-optimized 3DIC and 3D package design built in a multi-tier 45 nm process. Results obtained using high level models are shown in Table 4. The advantages of a 3D implementation are substantial.

Table 3. Performance requirements of a 3D SAR processor.

FPU Performance	1.8 TFLOPS
Memory Bandwidth	600 GBps
Memory Capacity	16+ GB

Table 4. Results of a design study investigating alternative implementations for a 3D SAR processor.

Implementation	Power	Area
COTS, using Cell Processors and XDR RAMs	743 W	48” x 48”
45 nm energy-optimized 2D design	260 W	32” x 32”
45 nm energy-optimized 3D design	130 W	3” x 3”

3. COMPUTER-AIDED DESIGN FLOW

The design-flow used for the FFT and Open RISC core was similar to any RTL-to-GDS flow using SoC Encounter from Cadence, but with some inter-tier constraints. After the partitioning stage, each of the steps is a multi-tier version of the usual EDA flow. During each stage, the feasibility of the design is checked and if not found feasible, another iteration is done to tweak the design. Examples of non-feasible characteristics

include overlapping cell placement, TSVs not aligned correctly between tiers, insufficient routing resources, etc.

The design-flow begins by swapping the FIFO memories in the RTL code with code including three banks of memory. A single-tier version was created for comparison, using two banks. The 3-tier design is then synthesized and put in to the RTL-to-GDS flow illustrated in Figure 2. The design is first partitioned into the number of tiers available (3 in our case) using K-Metis. Because K-Metis does not handle the size difference between the standard-cells and memories, the memories are removed from the netlist before partitioning, and added back in afterwards. After partitioning, the top module contains only three sub-modules, corresponding to each tier. No cells are moved between tiers after this point.

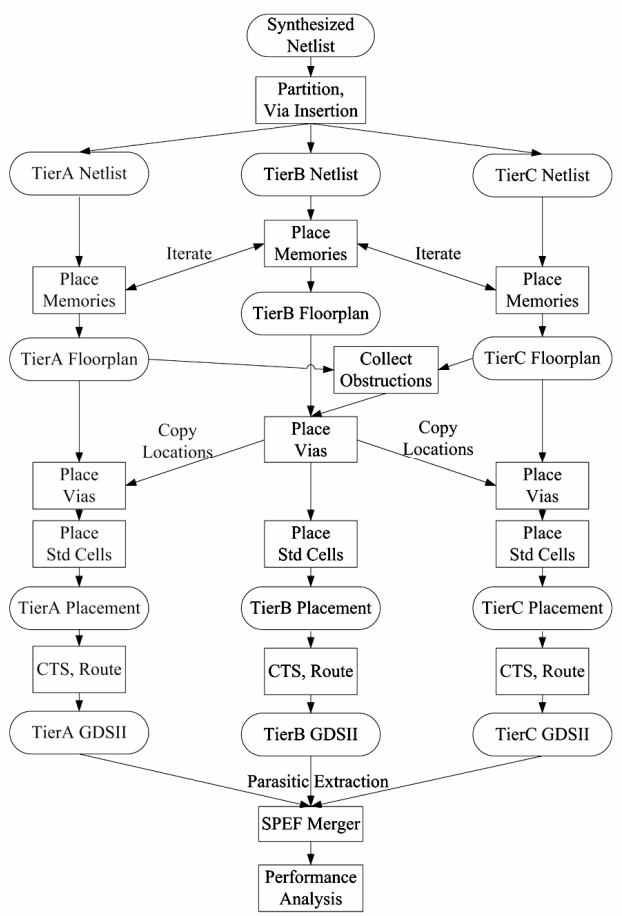


Figure 2. Physical 3D-IC Design Flow.

4. THERMAL MODELING

Thermal analysis is growing in importance for integrated circuits, because the main sources of failure in integrated circuits, electro-migration and gate-oxide breakdown, are both accelerated at high temperatures. For this reason, manufacturers typically will target a junction temperature of about 90-100 degrees C. 3D integration exacerbates this problem by increasing the power density of an integrated circuit, roughly multiplying it by the number of tiers and making it harder to ensure that the junction temperature remains below 100 °C. Under such constrained design conditions, accurate thermal analysis becomes very important to ensure that a design is meeting the temperature requirement.

The cooling difficulty is made worse by the fact that most of the silicon layers are essentially surrounded by thermal insulators, i.e. oxide. This fact increases the importance of accurate thermal impedance modeling. For example, Figure 3 shows a temperature map for one silicon tier in a dual core OpenRISC design. The temperature spikes are located at the clock buffers. Since clock buffers have high activity factors, this is not a surprising result. The increase in temperature leads to an increase in clock skew. If not correctly accounted for during timing design and verification, this could lead in turn to a timing failure in the produced part.

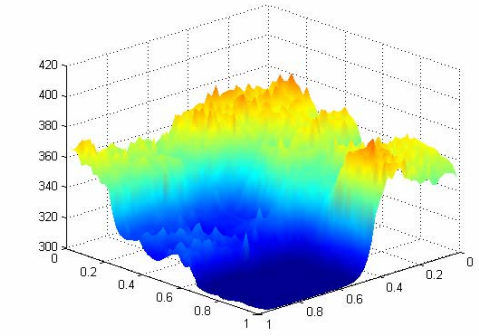


Figure 3. Temperature map for one layer of a 3-layer OpenRISC core, and as modeled in the CAD tool flow described in this paper. Temperatures are given in Kelvin. The temperature spikes are located at clock buffers.

Great advances have been made in recent years in the area of thermal analysis. Techniques such as the alternating direction implicit method and multi-grid approach [6] allow full-chip thermal simulation for planar ICs. These analysis methods are available in commercial tools. However, these tools usually assume isotropic thermal conductivity and use simple averaging approaches to calculate thermal conductivity. Because 3DICs place an insulator between all but one of the heat producing layers and the heat sink, details of the metallization and anisotropic thermal conductance will matter more. The impact of adding these two details to the thermal models is explored in the next two sub-sections.

4.1 Impact of Local Metalization

Current thermal modeling approaches ignore how the details of local metallization modify thermal conductivity. To illustrate the effect of adding metallization details at the transistor level, we simulated the simple model shown in Figure 4, which consists of a 17x9x65 grid of thermal elements. Each element models a block of material that is 10 μm long in the *x*- and *y*-directions. Each layer of elements in the *z*- direction corresponds directly with a layer in the physical process, and the elements are sized to match the thickness of the process layer they are modeling. The thermal simulations were performed in a nominal three-tier (three silicon layers) 45 nm SOI process.

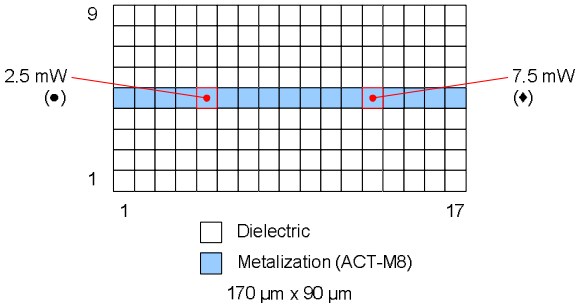


Figure 4. Physical 3D-IC Design Flow.

The model includes two heat sources: one 2.5 mW source and one 7.5 mW heat source. The total power flux of the simulation region is 65 W/cm^2. Lines of metallization have been laid down such that they run horizontally between the power sources. The metallization consists of solid from the active layer (ACT) through metal eight (M8). Solid fill was selected for all layers to ensure that we looked at the most extreme cases, with no vertical metallization.

Separate simulations are performed with the design (heat sources and metallization) on each tier. The tiers are: A (bottom), B (middle, inverted), C (top, inverted). For each tier, the simulations are run with and without metallization. For each of the six configurations of tier and metallization, four simulations are run. The heat sources begin at location x=5 and x=13, and are incrementally moved until they are located at x=9 and x=10 in the final simulation.

The results are summarized in Table 5. Tier A is closest to the heat-sink, while tier C is the farthest. Neglecting the horizontal metallization leads to a mis-prediction in temperature rise of over 1,000 °C. However, note that the mis-prediction is five times less on Tier A, which is the surrogate for the 2DIC result. Including details of horizontal metallization is significantly more important in 3DIC over 2DIC.

Table 5. Predicted temperature rises with and without horizontal metallization included in the thermal model. Tier A is closest to the heatsink.

Highest Temperature Rise	**Tier A**	**Tier B**	**Tier C**
No Metalization	200 °C	1,200 °C	2,000 °C
With Horizontal Metal	30 °C	113 °C	174 °C

4.2 Anisotropic Thermal Conductivity

State of the art full-chip thermal analysis, such as the technique used in [6] tends to break the chip down into a manageable set of homogeneous blocks with thermal conductivities in the x, y, and z directions. A conceptual model of this block is shown in Figure 5. The resistors in the model represent the thermal resistivity in each direction. The terminals represent the connections to adjacent blocks.

One of the most significant unsolved problems with this approach is how to determine the thermal conductivities. The common approach is to assume a homogeneous material, with the uniform thermal conductivity calculated as the weighted average of the thermal conductivities of the materials within each block. This approach ignores the potential directionality of conductivity due to metal routing.

Techniques have been developed to extract effective thermal conductivities for blocks containing a random arrangement of two materials [7-8]. We will extend the technique presented in [7] to accurately calculate effective conductivities when the composite material is known exactly, as is the case of an integrated circuit layout.

The thermal conductivity of an arbitrary block of material is governed by Fourier's Law of Heat Conduction:

$$q = -\kappa \nabla T$$

where q is the heat flux vector, κ is the thermal conductivity tensor and T is the temperature.

This can also be written in matrix form as:

$$\begin{bmatrix} q_x \\ q_y \\ q_z \end{bmatrix} = - \begin{bmatrix} \kappa_{xx} & \kappa_{xy} & \kappa_{xz} \\ \kappa_{xy} & \kappa_{yy} & \kappa_{yz} \\ \kappa_{xz} & \kappa_{yz} & \kappa_{zz} \end{bmatrix} \begin{bmatrix} \partial x / \partial T \\ \partial y / \partial T \\ \partial z / \partial T \end{bmatrix}$$

Homogeneous or isotropic materials will have all diagonal elements equal, with off-diagonal elements equal to zero. An orthotropic material will also have a diagonal matrix, but each term on the diagonal will be different. The model shown in Figure 5 can be used to accurately simulate an orthotropic material. However, such a model still ignores some of the potential directionality of heat flow. Fully anisotropic materials will have a non-diagonal, positive definite conductivity matrix. The largest eigenvalue of this matrix represents the maximum conductivity for the material, and its eigenvector represents the axis of maximum conductivity.

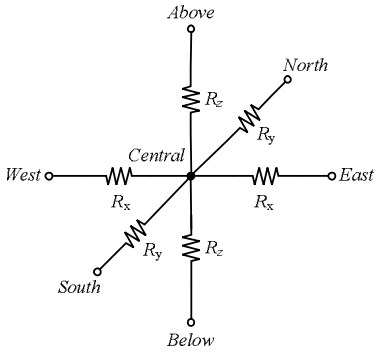


Figure 5: Common Thermal Block Model

4.2.1 Parameter Extraction

An experiment was performed using the metal layout shown in Figure 6. The materials in this example were chosen to be representative of a 45 nm CMOS process. Each block is assumed to be 65 nm wide and high and 130 nm thick and modeled using parameters obtained from [8-11].

The values κ_{xx} and κ_{xy} can be found by setting up a non-zero voltage (representing $\partial x/\partial T$) between the left and right edges of the block and a zero voltage between the top and bottom edges (representing $\partial y/\partial T$), as shown in Figure 7 for a 4x4

grid. The values q_x and q_y are the sum of the currents through the zero-volt sources on the top and right sides, respectively. It is important to use separate sources for each column of blocks, because a single source would create a direct current path between the left and right sides along the edge, introducing error into the result.

This approach leads to two equations and two unknowns, which can easily be solved. To determine κ_{yy}, a second simulation is required in which $\partial x/\partial T$ is zero and $\partial y/\partial T$ is non-zero. In a three dimensional problem, $\partial z/\partial T$ would be zero for the first two simulations, allowing the calculation of κ_{xz} and κ_{yz} with three equations and unknowns. A third simulation is necessary to find κ_{zz}, with both $\partial x/\partial T$ and $\partial y/\partial T$ set to zero.

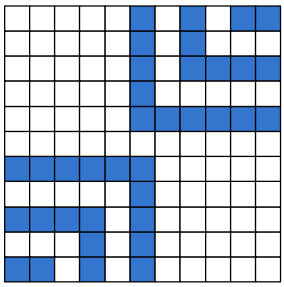


Figure 6. Layout modeled in this example. Metal in blue, Dielectric in white.

Simulation of the structure in Figure 6 yields a conductivity matrix with eigenvalues of 15.98 W/m°K and 0.42 W/m°K, with eigenvectors rotated 45° from the primary axes. The standard approach of assuming a weighted average yields a conductivity of 43.65 W/m°K in all directions, which is an overestimate of the maximum directional conductivity and a gross over-estimate of the orthogonal conductivity.

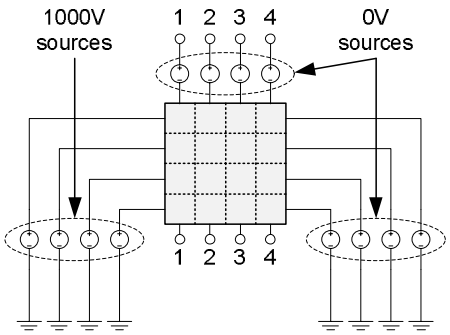


Figure 7. Full model including boundary conditions

4.3 Thermal Extraction and Modeling Flow

The ultimate goal of this work is to develop a thermal extraction and modeling flow that will provide sufficient fidelity for 3D IC design. The overall flow is shown in Figure 8. The layout is analyzed to determine a thermal resistance matrix for each macroblock as described above. This results in a high fidelity, anisotropic thermal resistance matrix. An important step in this process is determining the ports of interest in the overall model. For example, in circuits such as clock buffers, we want to know the temperature accurately and have accurate heat flux values. Thus, these might be treated as single ports. On the other hand, in logic blocks neither condition applies, and large blocks might be treated as single ports. Before sending the model to a simulator, an appropriate model reduction step is performed to condition it for the simulator.

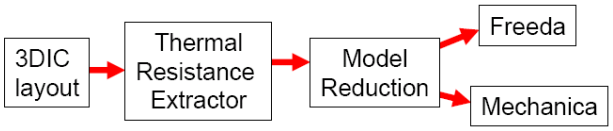


Figure 8. Overall Extraction and Modeling Thermal Flow.

4.4 Mechanica

The commercial thermal simulation program, Pro/ENGINEER Mechanica, was used to perform steady state thermal analyses. Mechanica is a general purpose p-version finite element program that was adapted to support this work. Specifically, new algorithms were implemented to extract the description of the 3D IC from the model reduction block shown in Figure 8, and to create a partitioning (or mesh) of the model that is composed entirely of hexahedral finite elements. With the p-version of the finite element method, the temperature within one element (or macroblock) can be approximated with up to a ninth degree polynomial in each direction, thus ensuring a high fidelity analysis solution. Steady state thermal analyses of 3D ICs were performed using orthotropic and anisotropic conductivities in the thermal macroblocks. Figure 9 shows a sample thermal modeling result.

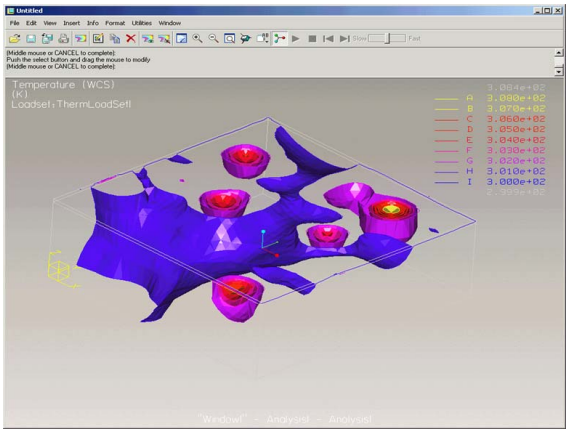


Figure 9. Sample temperature profile for a 3DIC.

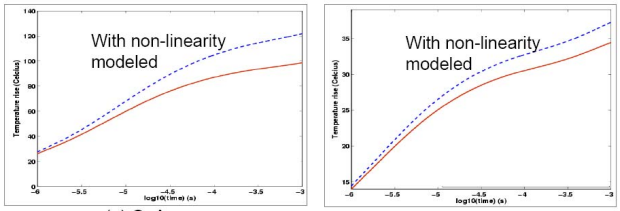
4.5 Future Work

As well as completing and demonstrating the flow above, planned future work includes investigating the modeling of non-linear materials and thermal transient modeling.

4.5.1 Non-linear materials

Many semiconductor materials are thermally non-linear which can cause errors in simulation. This is illustrated in Figure 10, which shows modeling results from two solid blocks of material, 400 µm cubed, with one face producing 0.40 W of heat. The plots show the temperature rise over ambient due to

this heat. It can be seen that a modeling error of up to 20% occurs if this effect is ignored.



(a) GaAs (b) Silicon

Figure 10. Effect of thermal non-linearity on predicted temperature rise in GaAs and Silicon.

4.5.2 Thermal transients

A straightforward way to extend this model to handle thermal transients is to modify the unit cell structure shown in Figure 5 with a structure like that shown in Figure 11. The capacitance represents the thermal capacity of the structure modeled by the unit cell.

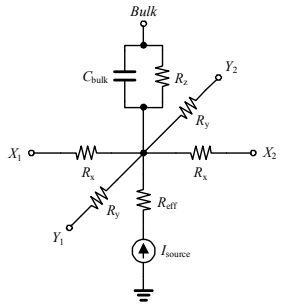


Figure 11. Transient simulations can be performed by including a heat capacity in the unit cell model.

5. CONCLUSIONS

3DIC technology with high density through silicon vias can lead to significant performance improvements in important applications. We presented a number of logic-on-logic applications in which critical path delay, and sometimes power, could be improved by around 30%. A memory-on-logic application was presented in which a better than six-fold improvement in power consumption was obtained through a large redesign that exploited 3DIC and 3D packaging technologies. Computer-aided design for these applications was carried out using a modified 2DIC CAD flow, based largely on commercial tools. Thermal design of 3DICs will require larger-scale, higher-fidelity modeling than for 2D ICs. In particular it will be important to model details of the thermal structure, especially anisotropic conductivity and detailed metal conductivity, in order to obtain sufficient fidelity.

6. ACKNOWLEDGMENTS

This project was funded by DARPA under contract FA8650-04-C-7127, managed by AFRL. We thank the DARPA program managers, Dan Radack and Michael Fritze, and the contract managers, Greg Creech and Steve Dooley. Additional funding was provided by the National Science Foundation under award 0643700.

7. REFERENCES

[1] W. Davis, J. Wilson, S. Mick, J. Xu, H. Hua, C. Mineo, A. Sule, M. Steer, and P.D. Franzon, "Demystifying 3D ICs: The Pros and Cons of Going Veritical," IEEE Design and Test of Computers, VOl. 222, No. 6, Nov-Dec, 2005, pp. 498-510.Bowman, M., Debray, S. K., and Peterson, L. L. 1993

[2] E.C. Oh, P.D. Franzon, "Design Considerations and benefits of Three-Dimensional Ternary Content Addressable Memory," Proc. IEEE CICC, Oct., 2007.

[3] H. Hua, "Design and Verification Methodology for Complex Three-Dimensional Digital Integrated Circuits," Ph.D. Dissertation, NC State University, 2006.

[4]] K. Schoenfliess, "Performance Analsysi of System-on-Chip Application of 3D Integrated Circuits," MS. Thesis, NC State University.

[5] H.P Hofstee, "Future Microprocessors and off-chip SOP Interconnect," in IEEE Trans. Advanced Packaging, Vol. 27, No. 2, May 2004, pp. 301-303.

[6] P. Li, et al., "IC thermal simulation and modeling via efficient multigrid-based approaches." IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 2006, Vol. 25, pp. 1763-1776.

[7] P. Phelan and R. Niemann, "Effective Thermal Conductivity of a Thin, Randomly Oriented Composite Material." Journal of Heat Transfer, s.l., ASME, 1998, Vol. 120, pp. 971-976.

[8] J. Wang, et al., "A new approach to modelling the effective thermal conductivity of heterogeneous materials." International Journal of Heat and Mass Transfer, 2006, Vol. 49, pp. 3075-3083.

[9] R. Hoofman, et al., "Challenges in the implementation of low-k dielectrics in the back-end of line." Microelectronic Engineering, 2005, Vol. 80, pp. 337-344.

[10] Z. Luo, et al., "High Performance Transistors Featured in an Aggressively Scaled 45nm Bulk CMOS Technology." 2007. pp. 16-17.

[11] E. Richard, et al., "Manufacturability and Speed Performance Demonstration of Porous ULK (k=2.5) for a 45nm CMOS Platform." 2007. pp. 178-17.