An Automated Design Flow for 3D Microarchitecture Evaluation*

Jason Cong[1] Ashok Jagannathan[1] Yuchun Ma[1,2] Glenn Reinman[1] Jie Wei[1] Yan Zhang[1]
[1]University of California, Los Angeles, California 90095, U. S. A.
[2]Tsinghua University, Beijing, P.R.China

Abstract - Although the emerging three-dimensional integration technology can significantly reduce interconnect delay, chip area, and power dissipation in nanometer technologies, its impact on overall system performance is still poorly understood due to the lack of tools and systematic flows to evaluate 3D microarchitectural designs. The contribution of this paper is the development of MEVA-3D, an automated physical design and architecture performance estimation flow for 3D architectural evaluation which includes 3D floorplanning, routing, interconnect pipelining and automated thermal via insertion, and associated die size, performance, and thermal modeling capabilities. We apply this flow to a simple, out-of-order superscalar microprocessor to evaluate the performance and thermal behavior in 2D and 3D designs, and demonstrate the value of MEVA-3D in providing quantitative evaluation results to guide 3D architecture designs. In particular, we show that it is feasible to manage thermal challenges with a combination of thermal vias and double-sided heat sinks, and report modest system performance gains in 3D designs for these simple test examples.

1. Introduction

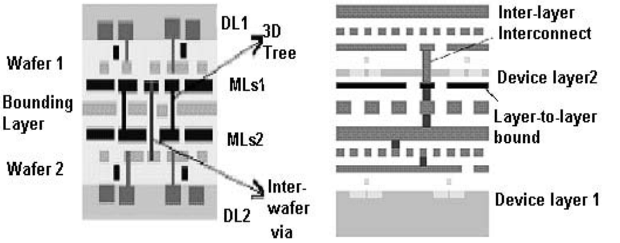
Due to the aggressive scaling of semiconductor technology, interconnect delays have become the dominant factor limiting system performance. Three-dimensional integration (3D ICs) has been proposed, such as [1, 4, 5, 13, 25, 28], to address this problem by reducing the wirelength, and therefore the interconnect delay, in future technologies. Moreover, reduced interconnection lengths in 3D ICs will help reduce power dissipation since the number of repeaters required for the wires can be reduced and the wire capacity load is reduced. However, thermal dissipation is one of the biggest concerns in 3D designs. The increased power density due to the decreased die area and the low conductivity inter-layer dielectrics (ILD) between the device layers can lead to thermal problems. For this reason, we need to evaluate 3D integration systematically, with regard to both performance and temperature.

Although a significant reduction of wirelength has been reported using 3D technology [13, 14, 15], its impact on the microarchitecture is still poorly understood due to the lack of tools that enable such an evaluation. While earlier work studied the impact of 3D IC on the memory subsystem [19] of the architecture or a specific implementation of an IA32 microprocessor [5], there is no existing flow that allows us to evaluate 3D implementations of architectures systematically and study them from both a performance and thermal perspective. We present an automated 3D microarchitecture evaluation flow (MEVA-3D) that can be used to compare and contrast 2D and 3D implementations of a given architecture. Specifically, we make the following contributions in this work:

First, we develop a floorplanner to optimize a microarchitecture for area, wirelength, performance, and temperature both for 2D and 3D integration. The performance is optimized by reducing the latency along critical loops in the architecture through interconnect pipelining at a given target frequency. Thermal evaluation is done using a resistive network model considering whitespace and thermal via insertion. Finally, we use our 3D router to do a detailed physical implementation.

Based on our automated evaluation flow, we study the impact of 3D integration on the performance and temperature of a processor model at 70nm technology and target frequencies ranging from 3GHz to 8GHz. For this simple architecture design driver, we can improve performance by around 11% at several frequencies by eliminating most of the wire latencies in 2D. We also show that the on-chip temperature can vary by up to 4.78x depending on what 3D IC fabrication technology is used, and we show that it is feasible to manage temperature challenges through a combination of thermal vias and double-sided heat sinks.

The remainder of this paper is organized as follows: Section 2 gives a brief overview of the different 3D fabrication technologies and the technology assumption we use in this paper; Section 3 introduces our architectural evaluation framework, including thermal evaluation and optimization, performance estimation, and the microarchitectural floorplanner with wire pipelining; Section 4 presents the baseline architecture which is evaluated by our framework and the performance and the temperature results of the baseline architecture. We conclude the paper and discuss the possible future research directions in Section 5.



(a) With flipping the top layer (b) Without flipping the top layer

Fig. 1 A 3D IC example with two device layers

2. Technology Background

3-D IC fabrication technologies include multi-chip module (MCM) packaging [1, 25], wafer bonding [4, 5, 13], solid- phase recrystallization [4], etc. Different fabrication technologies will greatly affect the circuit performance, manufacturing cost, on-chip temperature, etc. In this work we will use wafer bonding 3D IC technology, where each device layer is fabricated separately and then put together into one design. We will evaluate the same microprocessor architecture in two different kinds of wafer bonding strategies described in [5, 13], as shown in Fig. 1. Both technologies contain two device layers: in case (a), the top device layer is flipped upside down, and in case (b), the top device layer is oriented in the same direction as the bottom device layer. In case (a), the interlayer connect vias do not go through the device layers

This research is supported by the National Science Foundation under Grant CCR-0096383 and DARPA under Prime Contract DAAH01-03-C-R193 through CFDRC Corporation.

0-7803-9451-8/06/$20.00 ©2006 IEEE.

and therefore do not increase the chip size. Also in case (a), both silicon device layers are close to the boundary of the circuit, which will help the heat dissipation. However, case (a) cannot handle more than two device layers.

3. 3D Microarchitecture Evaluation Flow

Evaluating 3D architectures is a complex process that involves optimizing several parameters such as frequency, die area, performance, whitespace for thermal via insertion, and the temperature itself. An overview of our 3D Microarchitecture Evaluation Flow (MEVA-3D) with its components is shown in Fig. 2. The upper half of Fig. 2 shows the estimation part of our framework and the bottom half shows the validation part, which is used to verify the results generated by the estimation process.

The MEVA-3D flow takes as input: (1) a microarchitecture description in terms of the parameters of the blocks, such as the sizes of the structures, number of functional units etc.; (2) a target frequency for evaluating the microarchitecture; (3) a set of architecture-level critical paths whose latency will impact the performance of the microarchitecture, and a notion of performance sensitivity for each critical path; and (4) estimated power density numbers for the blocks in the processor. Notice that (3) and (4) are estimates of performance and power for the microarchitecture components that can drive a subsequent physical planning process.

In MEVA-3D, we propose an automated floorplanner that can be configured to optimize the floorplan for die area, performance, and temperature with consideration of interconnect pipelining. The performance estimation models are used to evaluate the impact of wire pipelining, and the power density estimates are used for the thermal calculation during floorplanning. The floorplanner is flexible enough to consider 2D and 3D placement of blocks, and allows one to configure the number of device layers for 3D integration. The output of the floorplanning engine consists of the locations of the blocks and their shapes, the total latency along the critical architectural paths including both blocks and wires, and the maximum on-chip temperature. Thermal via insertion and global routing can be employed to get the optimized thermal and routing profile. Once the exact block positions and wire latencies are known, this information is fed to our validation flow. A detailed cycle-accurate simulator that considers wire latency and power, and is coupled with power and thermal models, is used to validate the results from the MEVA-3D flow. In the sections that follow, we explain the components of our flow, mainly the performance optimization and the thermal evaluation methodology.

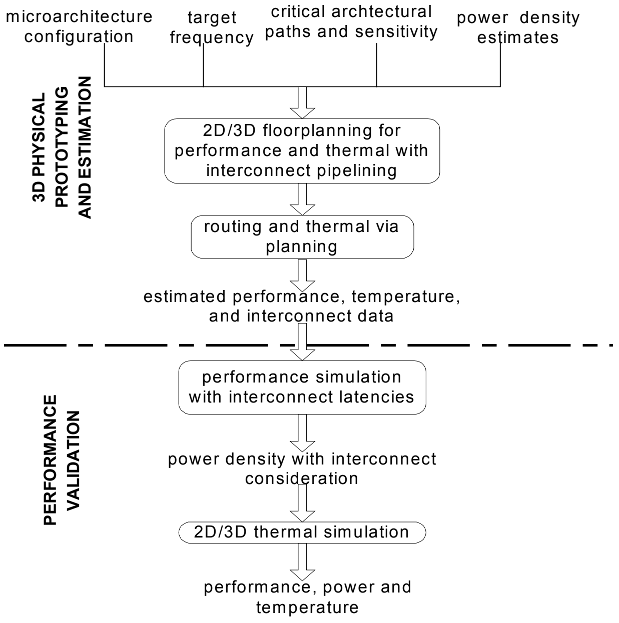

microarchitecture configuration
target frequency
critical archtectural paths and sensitivity
power density estimates
3D PHYSICAL PROTOTYPING AND ESTIMATION
2D/3D floorplanning for performance and thermal with interconnect pipelining
routing and thermal via planning
estimated performance, temperature, and interconnect data
PERFORMANCE VALIDATION
performance simulation with interconnect latencies
power density with interconnect consideration
2D/3D thermal simulation
performance, power and temperature

Fig. 2 Overview of MEVA-3D

3.1 Thermal Evaluation and Optimization

High on-chip temperatures will degrade timing, including both the gate and the interconnect delay, increase leakage power, and even cause logic failure. Therefore, temperature should be considered in any realistic 3-D microprocessor design. This section will describe the thermal evaluation tool and the automatic thermal via insertion tool that we use in the MEVA-3D flow.

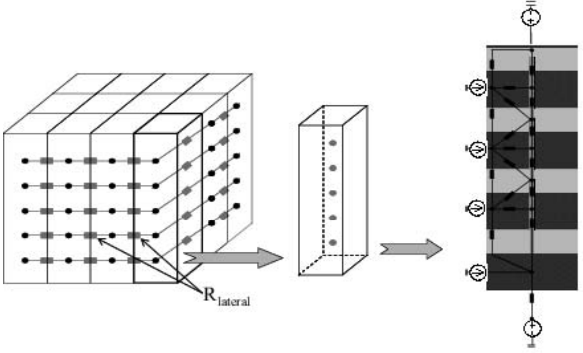
3.1.1 Resistive thermal network model

Considering both accuracy and runtime, we chose to make use of a compact, thermal resistive model proposed by Wilkerson [26], which explicitly models the 3-D circuit structure and the thermal vias. In [26], the circuit stack is first divided into tiles, each tile being the size of a thermal via pitch, as shown in Fig. 3(a). A tile stack is a vertical array of tiles, as shown in Fig. 3(b). A tile either contains one thermal via at the center, or no thermal via at all. A tile stack is modeled as a resistive network, as shown in Fig. 3(c). Fig. 3(c) shows the cases with two heat sinks. In cases with only one heat sink, one of the voltage sources can be taken away. The isothermal bases of room temperature are modeled as a voltage source. A current source is present at every node in the network to represent the heat sources. The tile stacks are connected by lateral resistances. The values of the resistances in the network are determined by a commercial FEM- based thermal simulation tool [3, 24] and it shows that our model has a small error within 5% [20].

3.1.2 Thermal via insertion

Chip cooling techniques can be divided into two groups. One of these is heat sink optimization, which tries to cool down the heat sinks through packaging level techniques such as fans and micro-channels. However, in 3-D designs, the poor thermal conducting ILD layers impede the internal heat dissipation from the heat sources to the heat sink. Even with the perfect heat sinks, the maximum on-chip temperature can still be very high. The other group of cooling technologies, which optimizes the internal heat dissipation paths, includes temperature-aware physical design tools [10, 11], thermal via insertion [11], and 3-D IC micro-channel techniques [15].

In this work, the temperature-aware floorplanning, thermal-aware routing and thermal via insertion techniques are applied to cool down the 3-D circuits. The temperature-aware floorplanning tool discussed in Section 3.3 tries to put the "hot" blocks close to the heat sinks to avoid hot spots. After floorplanning, we can further reduce the temperature by thermal via insertion. In [11], a simultaneous routing and thermal via planning method is proposed. Since the detailed netlist information is not available at the architectural level, we will concentrate more on the thermal via planning in the evaluation framework. A through-the-silicon via is a via that goes through a device layer. Under the current technology, through-the-silicon vias ($pitch \approx 5\mu m \times 5\mu m$) [16] are usually much larger than the metal wires. Through-the-silicon vias are very effective for heat dissipation and can decrease the maximum temperature over 50% [9]. When the signal vias are not sufficient to bring the chip temperature down to a satisfactory level, additional thermal vias can be inserted into the chip. The only purpose for thermal via insertion is for temperature reduction purposes —— the thermal vias have no wire connection. Experiments show that the temperature reduction will increase with the increase of the thermal via numbers, however, the temperature reduction will also gradually saturate with the increase of thermal via density. Therefore, we want to insert more thermal vias at hot places and spread the thermal vias whenever possible. The details of the thermal via planning algorithm can be found in [11].



(a) Tiles Stack Array (b) Single Tile Stack (c) Tile Stack Analysis

Fig. 3 Resistive thermal model for a 3D IC

3.2 Performance Estimation

Our approach to calculate the performance of the processor during the floorplanning process in Fig. 2 is similar to the method used in [18]. As the target frequency during floorplanning is fixed, we want to calculate the IPC degradation caused by the extra latency introduced by the interconnects in the layout. The work by Borch et al. [6] showed that the IPC of a microarchitecture depends on a set of critical processor loops, and extra latency along these loops can cause the IPC to degrade.

For each critical path, we develop IPC performance sensitivity models similar to the work by Sprangle and Carmean [23] which provides information about IPC degradation due to extra latency along the path. During floorplanning, we calculate the total latency of each critical path including the blocks and the wires, and determine the total number of cycles at the target frequency required to cover this path latency. Extra latency from the wires is used to compute the new IPC, and hence the performance of the processor for that floorplan.

3.3 Microarchitecture Floorplanning with Wire Pipelining

Traditional floorplanning optimizes the area and wirelength of the packing, but a minimal wirelength is not enough for performance optimization. Instead, the floorplanner should be used to effectively explore a large space of architectural configurations efficiently and accurately, measureing the trade-offs on each circuit path. This is particularly true for superscalar microarchitectures which can feature critical timing paths of varying importance —— minimizing overall wirelength may not be as important as minimizing critical loop wirelength.

The floorplanning problem that we investigate here considers several components in its objective function that are important tradeoffs in 3D architectures. Specifically, we consider the die area (footprint), the performance of the microarchitecture in *BIPS*, the maximum on-chip temperature, and the wirelength so that the power from the interconnects can be reduced. Formally, we define the problem as follows:

Given:

(1) target cycle time T_{cycle}

(2) clocking overhead $T_{overhead}$

(3) list of blocks in the microarchitecture with their area, dimensions and total logic delay

(4) set of critical microarchitectural paths with performance sensitivity models for the paths

(5) average power density estimates for the blocks.

Objective: Generate a floorplan which optimizes for the die area, performance, and maximum on-chip temperature.

The 3D thermal-driven floorplanner [10] used in this work is based on a simulated annealing framework with a 3D extension of the TCG representation [20]. We extended the floorplanner to consider the performance of the design instead of a wirelength optimization objective. Our cost function uses a weighted combination of area, performance, and temperature, and can be represented by

$$\cos t = w1 * \frac{1}{BIPS} + w2 * Area + w3 * Temp + w4 * Wire$$

where *BIPS* corresponds to the performance of the microarchitecture with that floorplan of the blocks, *Area* is the total area of the floorplan, and *Temp* corresponds to the maximum on-chip temperature calculated by the thermal model described in Section 3.1. The performance (*BIPS*) is calculated through the method presented in Section 3.2. The coefficients of *w1*, *w2*, *w3* and *w4* are used to control the different weight for each component. In our test evaluation, the performance component is given a high weight and will be optimized when the simulated annealing engine tries to minimize the cost function.

3.4 Performance Validation

Once we have finished the physical planning stage, we will have selected a 3D floorplan optimized for our performance and thermal estimation. Once we have the critical loop latencies and cycle time from this stage, we can input these, along with the architectural configuration, into our cycle-accurate simulation framework. We adapted the SimpleScalar 3.0 tool set [8], a suite of functional and timing simulation tools for the Alpha AXP ISA, for our simulation framework. This simulator allows execution-based simulation (including simulation down mispredicted branch paths) of a microarchitecture, with a wide range of customizable parameters. This framework gives performance statistics in instructions per cycle (IPC) that can be combined with the cycle time from the floorplanning stage to give a result in BIPS. This validation phase can provide feedback to the performance estimators in the floorplanning stage if the performance does not match the expected result.

4. Case Study for a Design Driver

In this section we present detailed evaluation results obtained for our design driver microarchitecture. This architecture, illustrated in Fig. 4, is an out-of-order microprocessor with detailed parameters shown in Table 1. We modified SimpleScalar to model this architecture and to parameterize the different critical path latencies found through the floorplanning process. To perform our evaluation, results were collected for 24 SPEC2000 [2] benchmarks.

As we consider automated floorplanning with performance and thermal optimization, it is necessary to estimate the area, wirelength, delay, and power associated with each block. We model

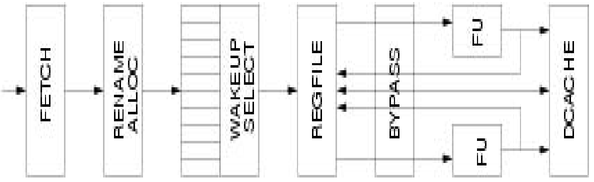


Fig. 4 Superscalar pipeline explored in this paper

the area and delay of the blocks using assumptions similar to Wilton and Jouppi [27] and Palacharla et al. [22] in which the authors analyzed the delay of many of the microprocessor logic blocks. The power numbers are modeled based on the models proposed by Brooks [7]. The delay of each block was derived by adding the delays along the critical path of the block. The aspect ratio for the blocks is also derived under these array layout assumptions. The power is calculated by adding the capacitances from the transistors as well as the wires inside these logic blocks.

Table 1. Baseline processor parameters

Instruction Cache	32KB, 32B/block, 2-way
Decode Width	8
ROB Size	128 entries
Issue Queue	32 entries
Issue Width	8
Register File	70 INT and 70 FP
Functional Units	Units 4 IntALU, 1 FPALU, 2 IntMult, 1 FPMult
Load/Store Queue	32 entries
L1Data Cache	16KB, 32B/block, 4-way, 2RW ports
Unified L2 cache	1MB, 64B/block, 8-way

4.1 Performance Impact of 3D Integration

In order to consider the impact of pipelining based on interconnect delays, we use the critical paths in Table 2 for this study. The area and delay of the blocks were derived based on [22, 27] for a futuristic 70*nm* process technology. Based on [17], we assume that the clock cycle overhead is 46*ps*, which corresponds to roughly 1.8FO4 (fan-out-of-four) for 70*nm* technology. Thus, for a 4GHz target cycle time, we set the useful time for computation as 204*ps* and use this to calculate the number of pipeline stages required to cover a given path delay. As we study the impact of semi-global interconnects in the microarchitecture, we assume that all of the signals are routed in the routing layers meant for semi-global wires. The delay of interconnects is derived using the IPEM models [12] which consider several optimizations such as wire sizing, buffer insertion and buffer sizing, etc.

To facilitate the insertion of repeaters, flip-flops, vias, etc., which are inevitable to achieve the required interconnect performance, we assume that 10% of each block's area is reserved around the block in the floorplan. Moreover, as the L2 cache occupies more than 50% of our die size, we allow the four L2 cache banks to be placed separately so that the floorplanner has more flexibility in packing the blocks. To study the impact of 3D integration on the performance of the microarchitecture, we generated the best performance results for the 2D and 3D cases by running the floorplanning engine 20 times and picking the best solution for each case. Fig. 5 shows that the impact of wire latency reduction due to 3D integration frequencies. Overall, we can see that a performance improvement of about 11% can be obtained by using a 3D design instead of a 2D design. Table 3 shows the detailed information for these test results. By using the 3D architecture, the die size can be shrunk to 53% of the die size in 2D design, and the wirelength can be reduced by about 46% at the same time. With the elimination of the extra latencies generated by the long wires along the critical paths, the performance measured in BIPS improves by around 11%. Table 4 shows the detailed information for the number of extra cycles caused by the wire delay in both 2D and 3D design. In 3D design, the long wires along the critical paths can be reduced, therefore, some of the extra cycles from wire can be eliminated. In Table 4, all the extra interconnect latencies are removed in 3D design for 3GHz frequency. Therefore we can see about 9.5% improvement in terms of performance. In 4GHz case, the critical path of ALU bypass has one extra cycle because the blocks along this path are spread out because of the limitation of 2D layout. But in 3D design, since the ALU units can be put on different layers to reduce the interconnection between them, the extra cycle for this loop is removed. Similarly, the extra cycles along the other loops are also reduced and the performance improvement from 2D design to 3D design is about 11.7%.

Table 2. Different critical paths considered during layout optimization

Wakeup Latency	Latency to wake up the dependent instruction
ALU Bypass	Latency of the bypass wires between the ALUs
DL1 Latency	Load latency though the L1 data cache
L2 Latency	Latency for access to L2 cache
MPLAT	Latency through the branch resolution path

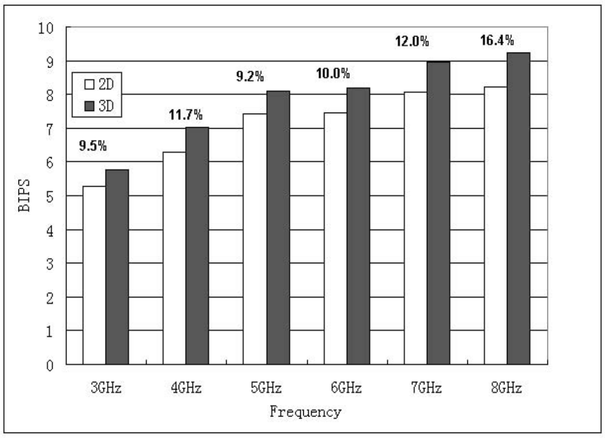


Fig. 5 Performance numbers for the microarchitecture with 2D and 3D layout at different target frequencies

Table 3. The results for different frequencies

Freq-uency		Area mm^2	Wire mm	BIPS	T(C) 1 sink	Via inserted		T(C) 2 sinks
						T(C)	Via Area %	
3G	2D	33.7	149	5.266	31.59	--	--	--
	3D	19.4	71.8	5.769	132.3	52.4	4.93	32.6
4G	2D	33.6	151	6.301	33.08	--	--	--
	3D	18.3	60.9	7.038	150.6	56.9	4.94	33.0
5G	2D	34.0	153	7.425	34.84	--	--	--
	3D	17.8	69.4	8.110	166.2	61	6.16	34.9
6G	2D	33.5	127	7.455	35.76	--	--	--
	3D	17.4	62.6	8.202	185.7	65.2	7.21	36.2
7G	2D	33.7	162	8.06	38.5	--	--	--
	3D	16.6	72.9	8.96	196.7	75.2	8.91	38.6
8G	2D	33.5	144	8.23	41.7	--	--	--
	3D	18.0	89.8	9.22	203.6	82.2	9.3	40.6
Average 3D/2D ratio		53%	48%	115%	478%	181%	--	100 %

To provide further insight, we show the performance simulation results for several benchmarks with the packing generated by our framework at a 6GHz target (at 6GHz frequency, the performance is the best in our test). In Fig. 6, we show ten test cases from the SPEC2000 [2] benchmarks. We can observe that the automatic floorplanner used in our work produces good quality floorplans, where blocks whose communication latency is critical for performance are placed close to one another in 2D as well as in 3D. The 3D layout eliminated the extra cycle from wires in the wakeup loop, and also removed some of the extra cycles on the branch misprediction resolution loop, the DL1 access loops, and the L2 access loops. Thus, the total improvement in performance comes to around 10.0% as shown in Fig. 5. Even though the 3D architecture could not remove all the extra cycles from the wire latency, it helped to reduce the latency along the most critical loops to gain good performance improvement. For the test cases shown in Fig. 6, each benchmark sees performance improvements in the range of 7% to 24%. On average, the performance improvement in Fig. 5 is about 10% —— a relatively large improvement given the simplicity of the microarchitecture, the limited die size, the small number of critical loops exposed in the floorplan optimization and simple re-use of 2D architecture components. More complex designs should see even more improvement from 3D technologies, especially when we consider the re-design of 3D components.

Table 4. The number extra cycles for critical paths group

	3G		4G		5G		6G		7G		8G	
	2D	3D	2D	3D	2D	3D	2D	3D	2D	3D	2D	3D
Wakeup	1	0	1	1	1	1	1	0	2	1	2	1
ALU	0	0	1	0	0	0	0	0	1	0	1	0
DL1	0	0	0	0	1	0	2	1	1	1	2	1
L2	2	0	2	1	3	1	5	1	5	2	5	1
MPLAT	1	0	1	0	2	1	2	1	2	1	5	2

4.2 Thermal Impact of 3D Integration

We also present thermal results corresponding to the best performance solutions presented in the previous section for both 2D and 3D. In order to calculate the power density of the blocks for thermal calculations, we use Wattch [7] models integrated with the simulator to calculate the dynamic power associated with each block at the specified target frequency. We add 30% of dynamic logic power as the additional power from the interconnects and associated repeaters. We chose this fraction based on observed interconnect power trends in recent studies [21]. We then assume that the leakage power is a fraction of the dynamic power, and add this fraction to obtain the total power dissipated by each block including the logic, wires, and leakage. For our experiments, we set the leakage to be 80% of the total dynamic power, as this corresponds to leakage contributing to 45% of the overall chip power. For reference, the leakage power of a Pentium 4 processor in 130nm technology is around 40% [21].* Based on the above method, we calculated the maximum power density at any point on the chip to be less than 2W/cm^2 for the 2D chip. While 3D integration usually helps to reduce interconnect power as well as the associated leakage power from the repeaters, we assume the same total power values for both the 2D and the 3D case in our thermal evaluation for easier comparison. Fig. 7 shows the maximum on-chip temperature obtained using the thermal simulator discussed in Section 3.1. The results are shown for the following four cases: (1) a 2D layout with a heat sink at the bottom (2D-HS1); (2) 3D integration in the case (b) technology in Fig. 1, with one heat sink at the bottom and without inserting any thermal vias (3D-HS1-no_via); (3) 3D integration in the case (b) technology in Fig. 1, with one heat sink at the bottom and thermal via insertion to reduce temperature (3D-HS1-via); and (4) 3D integration in the case (a) technology in Fig. 1, with two heat sinks, one at the top and one at the bottom (3D-HS2). As we expected, adding another device layer will generally cause a drastic temperature increase, and case (2) shows a temperature increase of over 4.78× on average. After thermal via insertion (case (3)), we can reduce the maximum on-chip temperature by an average of about 62%. And the area occupied by the thermal vias (Via Area %) is about 5% to 9% of the total chip area. However, we can see that the temperature is still quite high compared to case (4), under the wafer bonding technology with two heat sinks and the flipped top device layer. In case (4), both device layers are located very close to the heat sinks, which provides very good heat dissipation paths to the "hot" devices.

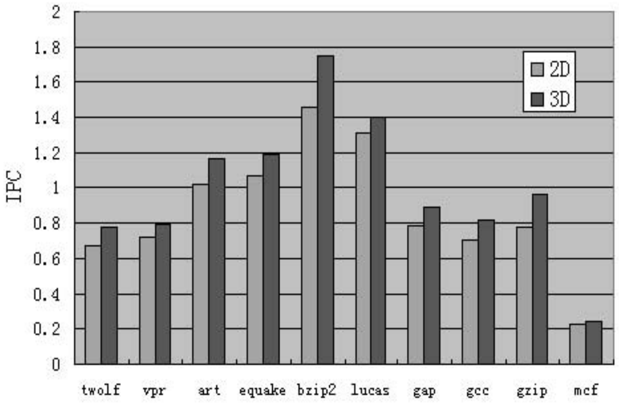


Fig. 6: Performance improvement for a 6GHz processor from 2D to 3D

5. Conclusions and Future Works

In this paper we presented an automatic evaluation flow named MEVA-3D for 3D architectures from 3D microarchitecture floorplanning, performance evaluation to thermal evaluation and thermal via planning. By using this flow, we can systematically evaluate the 3D architecture from both the performance side and the thermal side. Through one specific design driver, we see that 3D integration can help improve the performance by 10% with comparable maximum on-chip temperature. Since our automatic evaluation flow has the capability to handle different kinds of 3D architectures, there is great potential for an even further benefit from novel designs that can leverage 3D IC to improve individual blocks. To simulate the thermal effect dynamically, we are currently working to refine the power model. And the architecture components can be re-designed in 3D technologies, which can reduce the power consumptions and optimize the performance of the design.

[*] In this evaluation, since the design driver as both 2D and 3D designs have a similar temperature profile after thermal via insertion, we assume the leakage to be a fixed percentage of dynamic power.

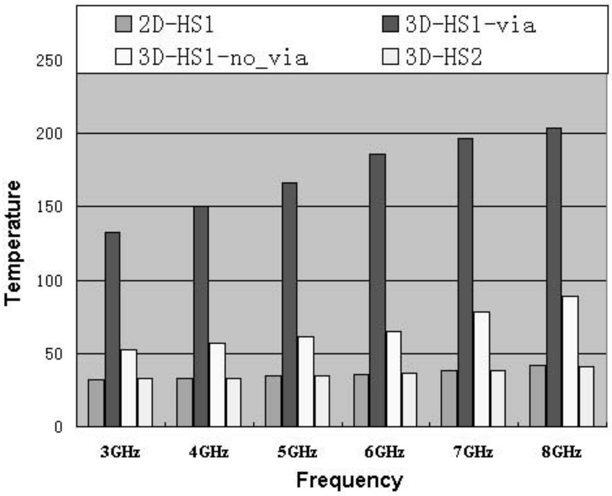


Fig. 7: Maximum on-chip temperature for different 2D and 3D integration for the best performance solution for the micro-architecture. HS denotes a heat sink, and the 3D integration allows the insertion of thermal vias to reduce temperature.

6. References

[1] *http://www.irvine-sensors.com/r_and_d.html#high*

[2] The Standard Performance Evaluation Corporation, 2000. http://www.spec.org.

[3] *CFD-ACE+ Module Manual*, 2002.

[4] K. Banerjee, S. Souri, P. Kapur, and K. Saraswat. 3-D ICs: A Novel Chip Design for Improving Deep-Submicrometer Interconnect Performance and Systems-on-Chip Integration, *Proc. of the IEEE*, 89(5):602–633, May 2001.

[5] B. Black, D. W. Nelson, C. Webb, and N. Samra. 3D Processing Technology and its Impact on IA32 Microprocessors. *Proc. Of ICCD,*pp.316-318,2004.

[6] Eric Borch, Eric Tune, Srilatha Manne, and Joel Emer. Loose Loops Sink Chips, *Proc. of 8th Internatonal Symposium on High-Performance Computer Architecture*, January 2002.

[7] D. Brooks, V. Tiwari, and M. Martonosi. Wattch: A Framework for Architectural-level Power Analysis and Optimizations. *Proc. of International Symposium on Computer Architecture*, June 2000.

[8] D. C. Burger and T. M. Austin. The SimpleScalar Tool Set, Version 2.0, Technical Report CS-TR-97-1342, University of Wisconsin, Madison, June 1997.

[9] T.-Y. Chiang, S. J. Souri, C. O. Chui, and K. C. Saraswat. Thermal Analysis of Heterogeneous 3-D ICs with Various Integration Scenarios, *IEEE International Electron Devices Meeting (IEDM) Technical Digest*, pp. 681–684, Dec. 2001.

[10] J. Cong, J.Wei, and Y. Zhang. A Thermal-Driven Floorplanning Algorithm for 3D ICs, *Proc. IEEE International Conference on Computer-Aided Design*, 2004.

[11] J. Cong and Y. Zhang. Thermal-Driven Multilevel Routing for 3-D ICs, *Proc. of the Asia South Pacific Design Automation Conference*, pp. 121–126, January 2005.

[12] Jason Cong and David Zhigang Pan. Interconnect Estimation And Planning For Deep Submicron Designs. *Proceedings of the 36th ACM/IEEE Conference on Design Automation*, pp. 507–510, 1999.

[13] S. Das, A. Chandrakasan, and R. Reif. Design Tools for 3-D Integrated Circuits, *Proc. Asia and South Pacific Design Automation Conf.*, pp. 53–56, January 2003.

[14] S. Das, A. Chandrakasan, and R. Reif. Timing, Energy, and Thermal Performance of Three-Dimensional Integrated Circuits, *Proc. of GLSVLSI*, 2004.

[15] Shamik Das. Design Automation and Analysis of Three-Dimentional Integrated Circuits, PhD thesis, Massachusetts Institute of Technology, May 2004.

[16] S. B. Horn. Vertically Integrated Sensor Arrays VISA (Invited). *Defense and Security Symposium*, 2004.

[17] M. S. Hrishikesh, K. Farkas, N. P. Jouppi, D. C. Burger, S. W. Keckler, and P. Sivakumar. The Optimal Logic Depth Per Pipeline Stage is 6 to 8 FO4 Inverter Delays. *Proc. of 29th International Symposium on Computer Architecture*, May 2002.

[18] Ashok Jagannathan, Hannah Honghua Yang, Kris Konigsfeld, Dan Milliron, Mosur Mohan, Michail Romesis, Glenn Reinman, and Jason Cong. Microarchitecture Evaluation with Floorplanning and Interconnect Pipelining, *Proc. of the Asia Pacific Design Automation Conference*, 2005.

[19] M. B. Kleiner, S. A. Kuhn, P. Ramm, and W. Weber. Performance and Improvement of the Memory Hierarchy of Risc-Systems by Application of 3-D Technology, *IEEE Trans. Comp. Packag, Manufact. Technol. B*, 19, 1996.

[20] J.-M. Lin and Y.-W. Chang. TCG: A Transitive Closure Graph Based Representation for Non-Slicing Floorplans, *Proc. of Design Automation Conference*, pp. 764–769, June. 2001.

[21] Nir Magen, Avinoam Kolodny, Uri Weiser, and Nachum Shamir. Interconnect-power Dissipation in a Microprocessor, *Proceedings of the 2004 International Workshop on System Level Interconnect Prediction*, pp. 7–13, 2004.

[22] Subbarao Palacharla, Norm Jouppi, and J. E. Smith. Complexity Effective Superscalar Processors, *Proc. International Symposium on Computer Architecture*, pp. 206–218, June 1997.

[23] Eric Sprangle and Doug Carmean. Increasing Processor Performance by Implementing Deeper Pipelines, *ISCA '02: Proceedings of the 29th Annual International Symposium on Computer Architecture*, pp. 25–34, 2002.

[24] Z. Q. Tan, M. Furmanczyk, M. Turowski, and A. Przekwas. CFD-Micromesh: A Fast Geometrical Modeling and Mesh Generation Tool for 3D Microsystem Simulations, *Int. Conf. MSM 2000*, pp. 712–715, March. 2000.

[25] Y. K. Tsui, S. W. R. Lee, J. S. Wu, J. K. Kim, and M. M. F Yuen. Three-Dimensional Packaging for Multi-chip Module with Through-the-Silicon Via Hole, *Electronics Packaging Technology, 2003 5th Conference*, pp. 1–7, 2003.

[26] P. Wilkerson, M. Furmanczyk, and M. Turowski. Compact Thermal Modeling Analysis for 3D Integrated Circuits. *11th International Conference Mixed Design of Integrated Circuits and Systems, Szczecin, Poland*, June. 2004.

[27] S. Wilton and N. Jouppi. CACTI: An Enhanced Cache Access and Cycle Time Model, *IEEE Journal of Solid-State Circuits*, May 1996.

[28] Y. Deng and W. Maly. Interconnect Characteristics of 2.5-D System Integration Scheme, *Proc. Int. Symp. On Physical Design*: pp.171-175, 2001.