The Impact of 3D Stacking on GPU-Accelerated Deep Neural Networks: an Experimental Study

William Wahby,* Thomas Sarvey,* Hardik Sharma,† Hadi Esmaeilzadeh,† and Muhannad S. Bakir*
*School of Electrical and Computer Engineering
Georgia Institute of Technology
{wwahby, tsarvey, msb}@gatech.edu
†School of Computer Science
Georgia Institute of Technology
{hsharma, hadi}@gatech.edu

***Abstract*—In this work, we present a two-tier air-cooled thermal testbed composed of an NVIDIA Tesla K40 GPU and a heater/thermometer top die. The top die has four independently-controllable heaters, which can emulate a wide range of components, ranging from low power memory to high-performance multi-core processor cores. The performance and temperature of the bottom-tier GPU on several deep neural network workloads is investigated as a function of increasing top-die power dissipation, and the implications for 3DIC cooling are discussed.**

***Index Terms*—Three-dimensional integrated circuits, deep neural networks, thermal management of electronics, thermal resistance.**

I. Introduction

Three dimensional integrated circuits (3DICs) are becoming an increasingly attractive option for system interconnection due to their potential to unlock ultra-high bandwidth [1]–[3]. Applications which require high bandwidth, such as machine learning [4], stand to benefit significantly from the heterogeneous 3D integration of high performance computing elements coupled with large quantities of memory. Thermal constraints complicate the design of such 3D systems, however, as the areal power density of a 3DIC can be much higher than the power density of the equivalent 2D system, making heat removal and thermal coupling significant challenges in 3D systems [5]–[7]. In order to begin to quantify the impact of thermal coupling on the performance of functional systems, we have developed a two-tier air-cooled 3D thermal testbed, shown in Fig. 1, composed of an NVIDIA Tesla K40 GPU [8] and a top die with resistive heaters, which can emulate a variety of different workloads.

II. Design and Assembly

The heater die (shown in Fig. 2) is composed of four serpentine platinum traces, each connected to two gold pads. Each quadrant of the die can be controlled and sensed separately, enabling the use of nonuniform power maps. The heater coils were fabricated via a lift-off process and are composed of a $0.2\mu m$-thick layer of platinum. The heater die was mounted back-to-back (B2B) with the GPU die. Since the heater die was stacked face up, the heater coils were covered with a layer of Kapton tape to electrically isolate them from the copper heat spreader, which slightly increased the thermal resistance of the stack. While a face-to-back (F2B) configuration would better reflect a typical 3D stacking scenario, the lack of clearance between the thermal die and the K40 board necessitated the B2B stacking approach. The heat spreader and heat sink were removed from the board, and the copper portion of the heat sink was milled down by approximately 0.5mm to accommodate the heater die, and an additional 0.5mm near the edges to accommodate the control/signal wires for the heaters, as can be seen in Figs. 3 to 5. Additionally, a portion of the aluminum board chassis was thinned down to allow the heater wires to exit the region immediately surrounding the GPU.

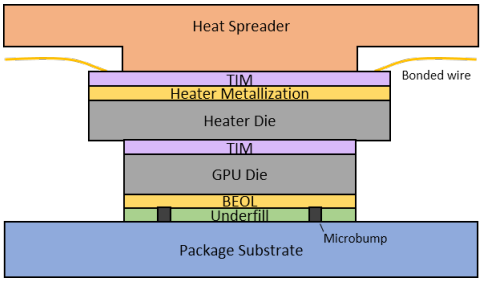

Fig. 1. Schematic of the air-cooled 3DIC thermal testbed.

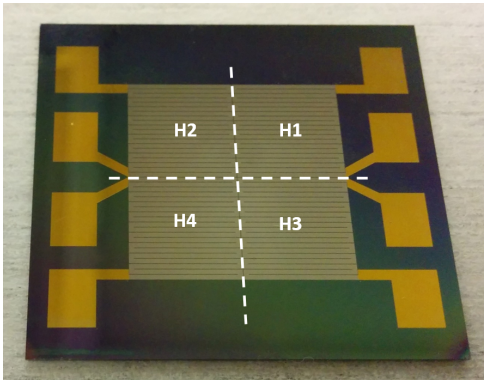

Fig. 2. Heater/thermometer die used to emulate a second tier logic or memory device. Each quadrant can be independently controlled.

To improve the thermal contact between the GPU, the heater, and the copper heat spreader, we used a thin layer of Arctic Silver 5 thermal interface material (TIM) at each interface. The resistance of each heater was measured over a range of temperatures in a Baxter Scientific Products DP-22 oven. As can be seen in Fig. 6, the heaters show a linear relationship between resistance and temperature. During operation, the heaters are driven at a constant power, and their

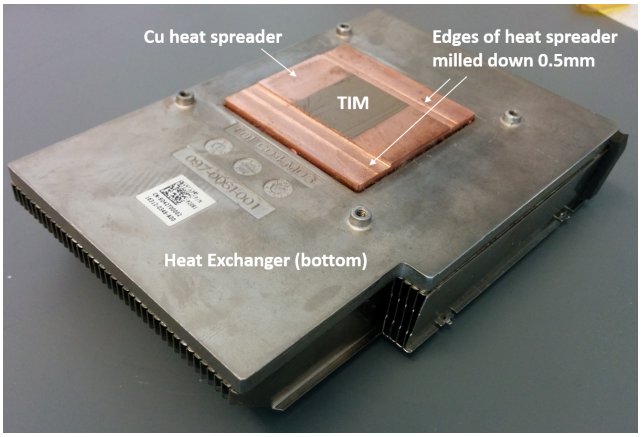

Fig. 3. NVIDIA Tesla K40 heat spreader with edges milled to accomodate heater/thermometer wires, and with thermal interface material applied to ensure efficient heat transfer.

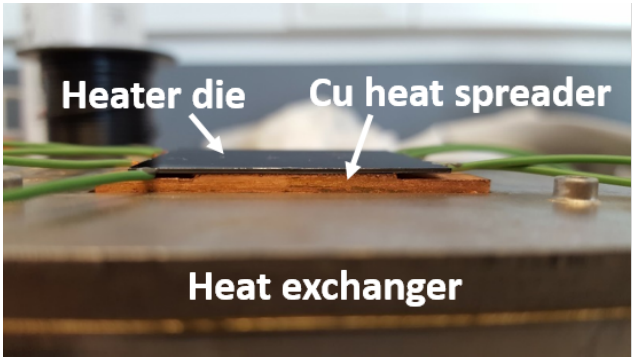

Fig. 4. Profile view of the modified heat spreader with top-tier heater/thermometer die attached. A small portion of the copper heat spreader was milled down to make room for the heater wires.

resistances are inferred from the driving voltages and currents. In order to validate the use of B2B stacking in the testbed, we simulated the thermal performance of a two-tier 3D stack, with a power density of 100W/cm^2 dissipated on the bottom tier, and 10W/cm^2 dissipated on the top tier. As can be seen in Fig. 7, the thermal difference between the two scenarios is very small, since the thermal conductivity of silicon is high. These results suggest that data from the B2B thermal testbed

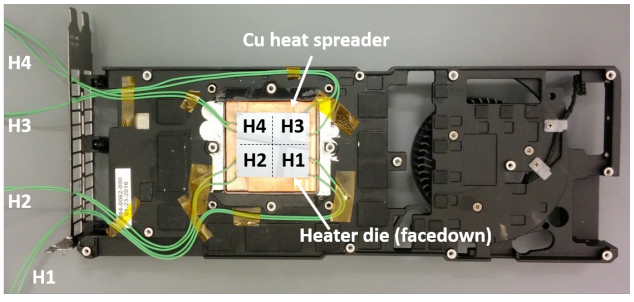

Fig. 5. NVIDIA Tesla K40 heat spreader with heater die attached. The heat spreader sits within a cutout in the aluminum chassis (black).

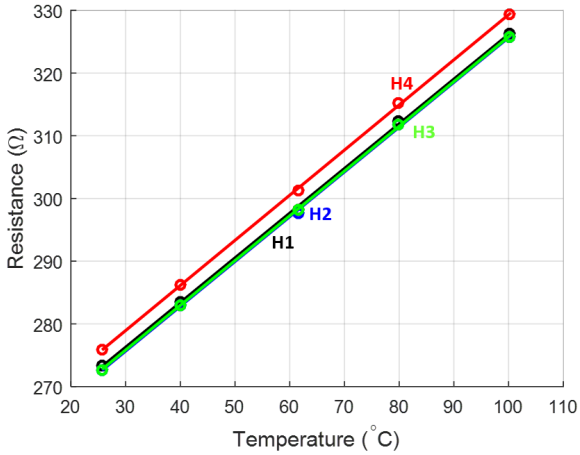

Fig. 6. Calibration measurements (circles) and best fits (lines) for the heater/thermometer structures on the top-tier heater die.

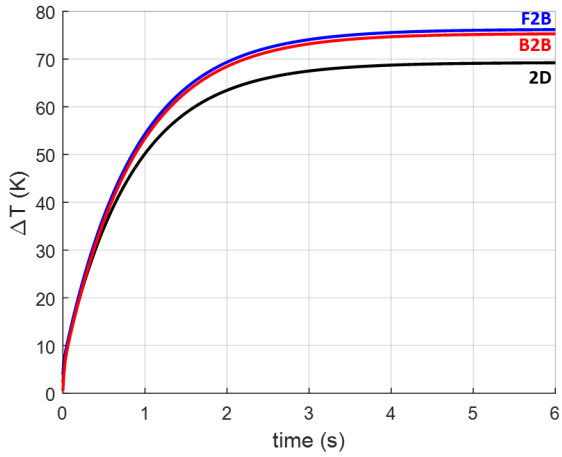

Fig. 7. Simulated thermal impact of face to back (F2B) vs. back to back (B2B) bonding of a two-tier 3D stack. The B2B thermal response very closely mirrors the F2B response, justifying the B2B approach used in the testbed.

can be used to make reasonable inferences about F2B systems.

III. Results and Discussion

We evaluated the two-tier thermal testbed with four deep neural networks (DNNs), detailed in Table I, which represent the state-of-the-art in artificial intelligence, recognition, and classification. Each DNN benchmark was run 25 times back to back to allow the GPU time to reach a steady-state condition under load, and the average GPU temperature, power consumption, and computation time were recorded. After each set of 25 runs, the system was kept idle for 5 minutes to allow time for the GPU to return to a baseline temperature after which the next benchmark was run 25 times back to back. This process was repeated for each benchmark with top-die power dissipations of 0W, 16W, 24W, 30W, and 40W. Each time the top die power dissipation was changed, the GPU was kept idle for 5 minutes to reach a steady-state temperature. After running the top die at 24W, the resistance of heaters 2 and 4 dropped to zero, due to a short caused by a small gap in

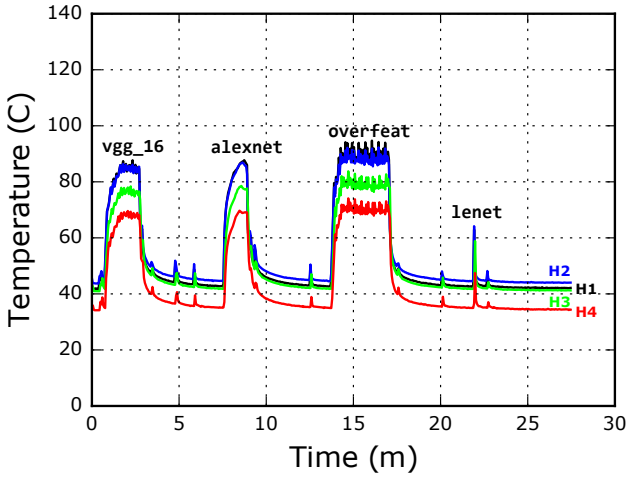

Fig. 8. Measured temperature of the top-tier die with the GPU running various machine learning workloads, and with the top die dissipating 0W.

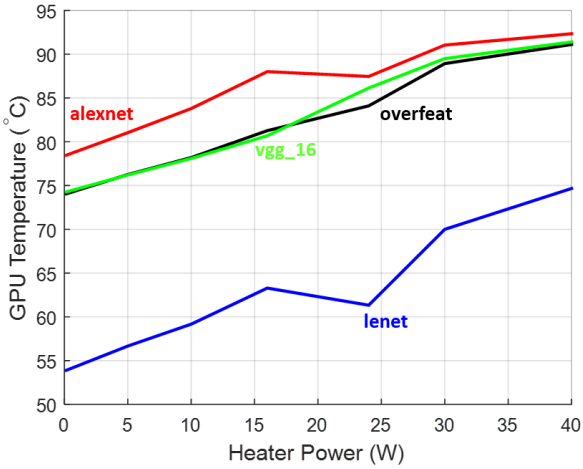

Fig. 9. Impact on GPU temperature as top-die power is increased.

TABLE I
BENCHMARK OVERVIEW

Network	Dataset	Domain	Model Size	MACCs
LeNet	MNIST	Digit recognition	0.8 MB	2M
AlexNet	ImageNet	Detect/classify	116.3 MB	736M
Overfeat	ImageNet	Detect/classify	278.3 MB	2,797M
VGG-16	ImageNet	Detect/classify	3.3.9 MB	16,361M

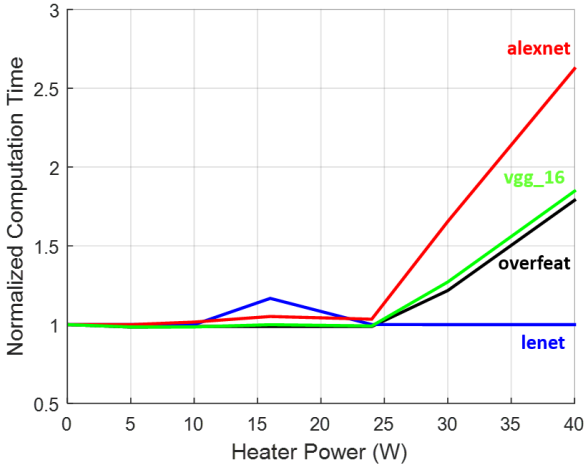

Fig. 10. Impact of heater power on GPU computation time. Each curve is normalized to its value at a top-die power dissipation of 0W.

the electrical isolation. To approximately compensate for the loss of heaters 2 and 4, heaters 1 and 3 were run at twice the power density for the 30W and 40W runs. The resistance of each heater on the top die was sampled every 3.3 seconds to determine the dynamic top-die temperatures.

In Fig. 8, the temperature of each heater on the top die is shown as a function of time for one complete test run encompassing all four benchmarks, with the top-die power dissipation set to 0W. The beginning and end of each workload are clearly visible as the die temperature rapidly increases to a steady state under load, then decays to its idle steady state. The variation in heater temperature is attributed to imperfect contact between the heat spreader and the heater/GPU stack.

In Fig. 9, the average GPU temperature during each workload is shown as a function of top-die power dissipation. As the top-die power dissipation increases, the average GPU temperature measured during each workload tends to increase, and the temperature during the AlexNet, Overfeat, and VGG-16 workloads exceeds 85°C at a top-die power dissipation of approximately 30W. The GPU remains relatively cool during the LeNet workload, as it has a much smaller computational footprint than the others, and does not fully stress the GPU. As shown in Fig. 10, the time required for each workload remains relatively flat until 30W, at which point the larger workloads begin to overpower the heat sink, and the GPU begins limiting its performance to avoid exceeding its thermal limits.

In Fig. 11, the average GPU power consumption is shown for each workload as a function of top-die power dissipation. GPU power increases for each of the large workloads (AlexNet, Overfeat, and VGG-16) up to a top-die power dissipation of approximately 24W, due in part to increased transistor leakage. Above 24W, the GPU power consumption drops sharply for each of the large workloads. As can be seen in Fig. 10, the GPU appears to limit its performance in order to remain within its thermal envelope, as the average computation time for each benchmark stays roughly constant until the average GPU temperature approaches 90°C, at which point the computation time dramatically increases. While the average GPU power decreases at high top-die power dissipations, the computation energy increases significantly, due to the increase in computation time, as seen in Fig. 12.

The temperature measured at heater 1 on the top die for each experimental condition is shown in Fig. 13. During the 30W and 40W tests the maximum temperature of heater 1 increases to 110°C. This high temperature can be attributed to the poor thermal transfer between heater 1 and the heat sink,

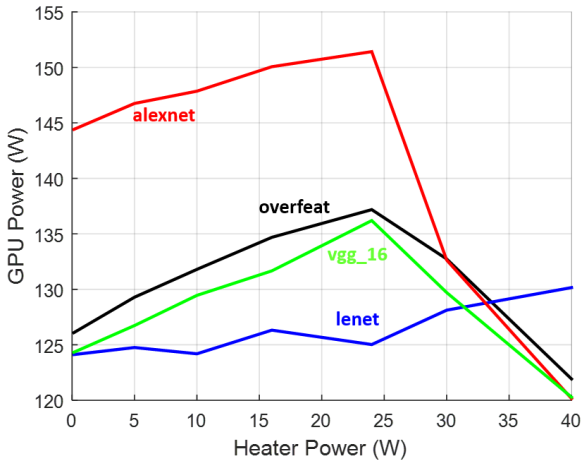

Fig. 11. Impact on GPU power dissipation as top-die power is increased.

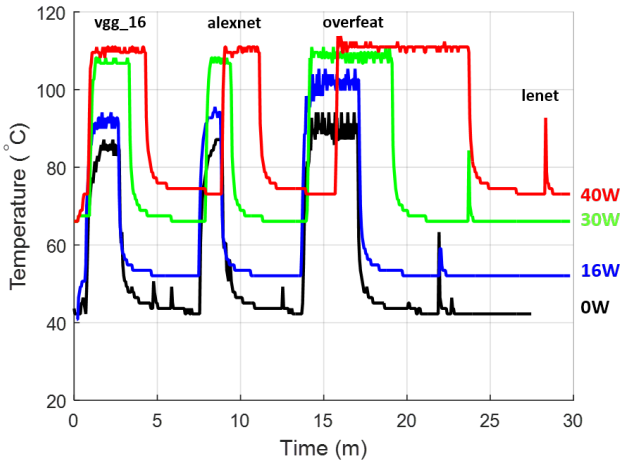

Fig. 13. Measured temperature of heater 1 during the benchmark suite, for a range of top-tier power dissipations.

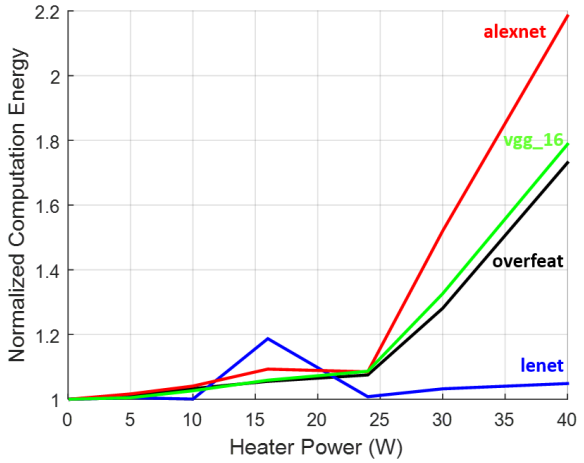

Fig. 12. Impact of heater power on GPU computation energy. Each curve is normalized to its value at a top-die power dissipation of 0W.

as shown in Fig. 8, and to the asymmetric power maps used for the 30W and 40W tests, during which only heaters 1 and 3 were used due to the failure of heaters 2 and 4.

IV. Conclusions

GPU-accelerated deep neural networks could benefit greatly from the high bandwidth and low latency enabled by 3D integration, as DNNs require large sets of model parameters to be fed to the cores of the GPU, but thermal limits could offset the benefits of such integration. In order to explore the impact of 3D stacking on DNN computational performance, we have developed and characterized an air-cooled thermal testbed for the investigation of the impact of thermal crosstalk and cooling limits on the performance of high performance 3DICs. The thermal testbed was used to emulate a two tier GPU-based 3D stack with a thermal die on the top tier to emulate stacked memory or logic. The GPU operating temperature increased steadily with top-die power dissipation, and once the average GPU temperature approached 90°C (at 30W top-die power dissipation), the GPU appears to limit its performance to avoid exceeding its thermal limits. In the worst case, we observed a 2.6X increase in computation time, and a 2.2X increase in computation energy, and we expect higher top-tier power dissipations to yield worse performance/efficiency degradation. These results suggest that aggressive cooling techniques may have a significant impact on the viability of high performance 3DICs, especially for DNN workloads.

Acknowledgments

The authors gratefully acknowledge the support of the SRC Educational Alliance Intel Foundation Fellowship, and NVIDIA Corporation for the donation of the Tesla K40 GPU used in this work.

References

[1] D. H. Woo, N. H. Seong, D. L. Lewis *et al.*, "An optimized 3D-stacked memory architecture by exploiting excessive, high-density TSV bandwidth," in *HPCA - 16 2010 The Sixteenth International Symposium on High-Performance Computer Architecture*, Jan 2010, pp. 1–12.

[2] G. H. Loh, "3D-Stacked Memory Architectures for Multi-core Processors," in *Proceedings of the 35th Annual International Symposium on Computer Architecture*. Washington, DC, USA: IEEE Computer Society, 2008, pp. 453–464.

[3] J. T. Pawlowski, "Hybrid Memory Cube (HMC)," in *Hot Chips: A Symposium on High Performance Chips*. Stanford, CA, USA: IEEE Technical Committee on Microprocessors and Microcomputers, in cooperation with ACM SIGARCH., 2011.

[4] B. Catanzaro, N. Sundaram, and K. Keutzer, "Fast support vector machine training and classification on graphics processors," in *Proceedings of the 25th International Conference on Machine Learning*, ser. ICML '08. New York, NY, USA: ACM, 2008, pp. 104–111.

[5] J. H. Lau and T. G. Yue, "Thermal management of 3D IC integration with TSV (through silicon via)," in *2009 59th Electronic Components and Technology Conference*, May 2009, pp. 635–640.

[6] H. C. Chien, J. H. Lau, Y. L. Chao *et al.*, "Thermal evaluation and analyses of 3D IC integration SiP with TSVs for network system applications," in *2012 IEEE 62nd Electronic Components and Technology Conference*, May 2012, pp. 1866–1873.

[7] W. Wahby, L. Zheng, Y. Zhang *et al.*, "A simulation tool for rapid investigation of trends in 3-DIC performance and power consumption," *Components, Packaging and Manufacturing Technology, IEEE Transactions on*, vol. 6, no. 2, pp. 192–199, Feb 2016.

[8] *Tesla K40 GPU Active Accelerator*, NVIDIA, November 2013.