Associative Processor Thermally Enables 3-D Integration of Processing and Memory

Leonid Yavits*, Amir Morad*, Ran Ginosar* and Eby G. Friedman**

*Department of Electrical Engineering, Technion - Israel Institute of Technology, Haifa, Israel

**Department of Electrical and Computer Engineering, University of Rochester, Rochester, New York, USA

Abstract— 3-D integration of conventional massively parallel processors is challenging due to high temperatures and hotspots. An Associative Processor (AP) is a viable candidate for such applications as it exhibits close to uniform thermal distribution with lower temperatures and fewer hot spots. Performance and thermal analysis supported by simulation confirm that associative processing enables 3-D integration of multilayer processing and memory cubes.

Index Terms — 3-D integration, SIMD, Associative Processor, Thermal Analysis

1 Introduction

Machine learning, data mining, network routing, search engines, and other big data applications can be significantly sped up by massively parallel machines, such as GPUs [7]. However, data transfer between the processing units (PUs) and on-chip memory hierarchies, as well as the off-chip memory bandwidth requirements limit the performance of massively parallel architectures [3] [27] [32], as illustrated in Fig. 1(a).

Three-dimensional (3-D) integration might be a natural step in high performance parallel processor evolution [32]. 3-D integrated circuits overcome both the on-chip and off-chip bandwidth limitations by bringing the memory much closer to the PUs, by stacking DRAM above the processors [11], as shown in Fig. 1(b). Additionally, multiple layers of massively parallel processors may be stacked, reducing the inter-PU communication latency and power, and facilitating closely-coupled parallelism. Such multilayer processing structures may be further enhanced by stacking multilayer DRAM memories within the same 3-D cube (see Fig. 1(c)).

When operating at high data rates, arrays of PUs are highly active, resulting in extreme power densities and hotspots [19], creating additional design constraints such as heat dissipation, power delivery, and excessive leakage current [20]. Unfortunately, 3-D integration cannot accommodate these constraints. High power densities and high temperatures adversely affect the performance of 3-D circuits [12] [22]. Although a variety of temperature-aware floorplan optimization techniques have been proposed [1] [9] [19] [30], thermal considerations remain a major concern in 3-D circuits. For example, placing DRAM above the PUs may be problematic if the temperature rises above the DRAM operational range (85℃-95℃ [15] [16] [17]). Such a scenario is described in [19]. Thus, a conventional high performance massively parallel computing architecture is not well suited for 3-D integration.

Associative Processors (AP) [18] [26] [35] may offer a viable alternative to conventional massively parallel PU arrays, such as GPUs. The AP, comprising a modified Content Addressable Memory (CAM), facilitates data processing in addition to content addressable and randomly accessible data storage. In essence, the computations are spread over the entire physical area, reducing thermal gradients and the overall temperature, and eliminating hotspots.

In this work, we study the thermal behavior of a 3-D integrated system of multiple AP dies (Fig. 1(c)). We show that an AP stack delivers close to uniform thermal density and low peak temperature, which enable stacking DRAM within and above the processing layers.

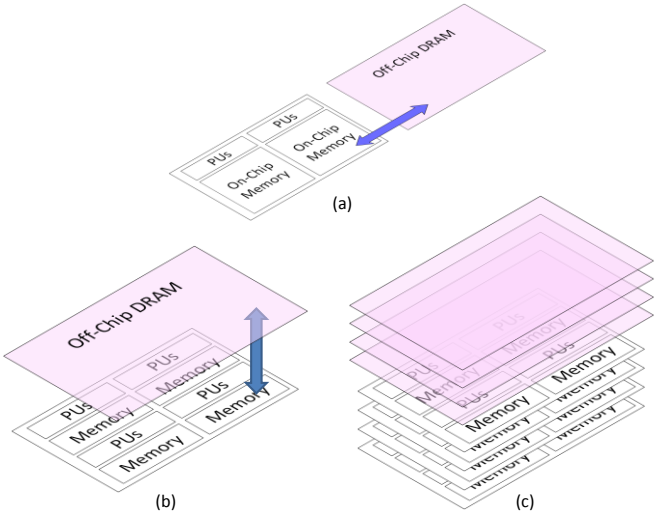

Fig. 1. Architecture evolution: (a) 2D PU array with on-chip and off-chip memory, (b) stacked PU array and DRAM dies, (c) 3-D multilayer PU/AP with multilayer DRAM

The rest of this paper is organized as follows. In Section 2, the experimental setup is described. The results of the thermal analysis of AP are presented in Section 3, while some conclusions are offered in Section 4.

2 Experimental Setup

The objective of the thermal analysis is to estimate the thermal distribution and peak temperatures (hot spots) within a 3-D AP. We perform the thermal analysis using the HotSpot simulator, a tool for architectural level thermal modeling [19] [24] [34].

Power trace and execution times required for HotSpot simulation are supplied by a cycle-accurate AP simulator [26]. In this section we discuss the workloads, describe the simulation methodology, and present the results.

2.1 Workloads

The following workloads have been selected for performance and power consumption simulation:

- N-option pairs Black-Scholes option pricing (BSC)
- N-point Fast Fourier Transform (FFT)
- Dense Matrix Multiplication of two $\sqrt{N}\times\sqrt{N}$ matrices (DMM)

where N is the data set size, for simplicity scaled to the AP size. The workload selection is based on the study by G. Almási *et al.* [2] who showed that scientific kernels should be used in evaluating the performance of massively parallel in-memory processors.

2.2 Simulator

We simulate the AP using a cycle-accurate simulator [26]. For FFT, we use an optimized parallel implementation, as outlined in [29]. For BSC, we used a direct implementation optimized for associative processing, based on the formulation described in [8]. Matrix multiplication uses the AP compare and arithmetic capabilities to match the input matrix element pairs and perform the multiplication. The singleton products are summed by the reduction tree [26].

The first step of AP programing is mapping the workloads on the associative processing array. For DMM, each pair of input matrix elements is processed by a single PU. For FFT, each butterfly is implemented by a single PU. For BSC, a single PU handles a single call option of a single security at a single strike price and single expiration time. At the next step, we break each fine-grain data thread into a series of arithmetic and data communication operations, and manually allocate temporary storage. At the last step, each arithmetic and communication operation is converted into a series of compares, writes and data moves. A description of these computations on AP is provided in [26]. Simulation times are listed in TABLE 1.

For the power simulation, our simulator [26] follows the methodology of SimpleScalar [4], maintains track of which PUs are active during execution, and records the total energy consumed by each PU for a workload. The simulator employs parameterized power models to estimate the energy consumed by each operation of each PU.

2.3 Results

We simulate the speedup and power per workload for 12 different values of N, ranging from 2^{12} to 2^{22}. In all cases, the PU size is 256 bits. The simulated speedup as a function of AP chip area is presented in Fig. 2(a). The DMM uses the reduction tree as an accelerator. BSC is a highly parallel workload. Hence DMM and BSC achieve higher speedup than FFT.

TABLE 1
Data Set Sizes and Simulation Times

Workload	Date Set Size	Simulation Time
BSC	$2^{12} \div 2^{22}$	11 sec ÷ 4 hours
FFT	$2^{12} \div 2^{22}$	10 sec ÷ 8 hours
DMM	$2^{12} \div 2^{22}$	7 sec ÷ 38 hours

Simulations performed on Intel® Core™2 Quad CPU Q8400 with 8GB RAM

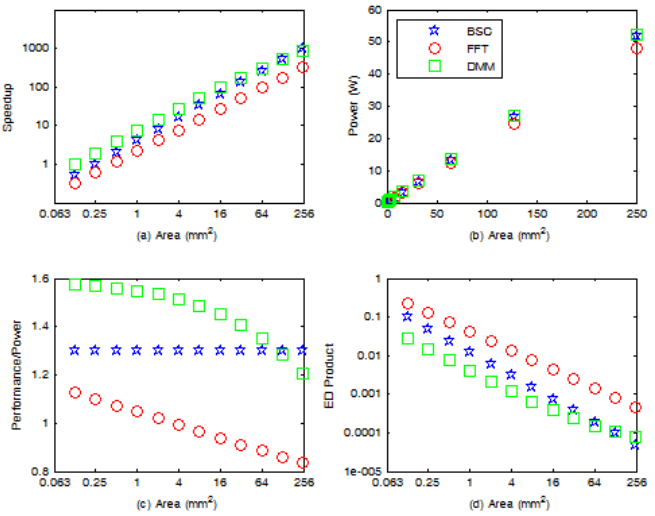

Fig. 2. Simulation results: (a) speedup, (b) power, (c) performance / power ratio, (d) energy × delay.

The power consumption results are presented in Fig. 2(b). All workloads consume approximately the same order of magnitude of power (hence we use a linear rather than log-log scale). This behavior occurs since all workloads use mostly identical associative primitives (compare and write). Although DMM uses a relatively power-hungry reduction tree, the reduction time is negligible as compared to the total runtime of the associative operations. The performance/power ratio and Energy × Delay (ED) product are shown, respectively, in Fig. 2(c) and Fig. 2(d). Among the workloads, DMM exhibits the best performance/power and ED, due to the accelerated reduction operation.

3 Results of 3-D Thermal Analysis

The floorplan of the AP used for HotSpot simulation is shown in Fig. 3. Due to the complexity restrictions of HotSpot, we simulate only a 4×4 mm AP die, which is divided into 1024 identical blocks (Fig. 3(a)). Each AP block features an associative processing array comprising 256 PUs of 256 bits each, a 256-bit TAG register, and 256-bit KEY and MASK registers (Fig. 3 (b)). The total number of PUs in the AP is 2^{18}. The thermal model of the 3-D AP is presented in Fig. 4. Four silicon layers (each containing the AP of Fig. 3(a)) are stacked over a Thermal Interface Material (TIM) layer, a Heat Spreader (HSP) layer, and a heat sink [11]. Following [19], ambient temperature of 46°C is

used, being typical for high performance servers.

The results of HotSpot simulation are presented in Fig. 5, which shows the thermal map of the AP at the top silicon layer of the 3-D stack. The peak temperature of this layer is 56℃. Note the contrast with the temperatures in a conventional GPU which may reach 105℃ [19]. The hottest region of the AP is located at the center, as expected, since during arithmetic operations, the active (switching) elements are uniformly distributed across the AP. The difference between the highest and lowest temperature in the AP is around 1.2℃. Clearly, the thermal distribution is close to uniform.

The thermal gradients of the four silicon layers of the AP along the T-Cut section (cf. Fig. 3(a)) are presented in Fig. 6. The peak temperature of the AP located in the upper silicon layer is below the maximum operating temperature of DRAM (85℃-95℃) [15] [16] [17], which enables 3-D DRAM integration above the AP layers. For comparison, a peak temperature of a GPU processor [19] is shown in Fig. 6.

Increasing the AP size to 4M PUs and 250 mm^2 produces a peak temperature of 85℃ and temperature gradient of 5.3℃ accros the die. Hence, a multilayer processor and memory cube is possible even for large die sizes. The thermal simulation results are summarized in TABLE 2.

TABLE 2
THERMAL SIMULATION OF A FOUR-DIE 3-D ASSOCIATIVE PROCESSOR

AP die size	AP die area	AP die power	Peak Temp	Temp Span
256K PU	16 mm^2	3 Watt	56℃	1.2℃
4M PU	250 mm^2	55 Watt	85℃	5.3℃

Temperatures in the upper silicon layer (layer 1)

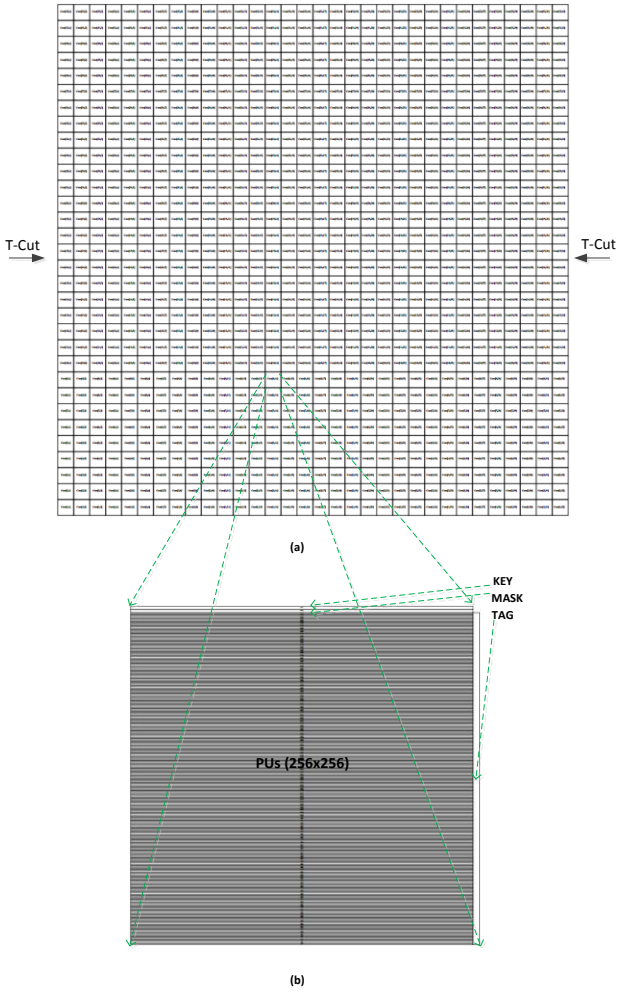

Fig. 3. AP floorplan for HotSpot thermal modeling

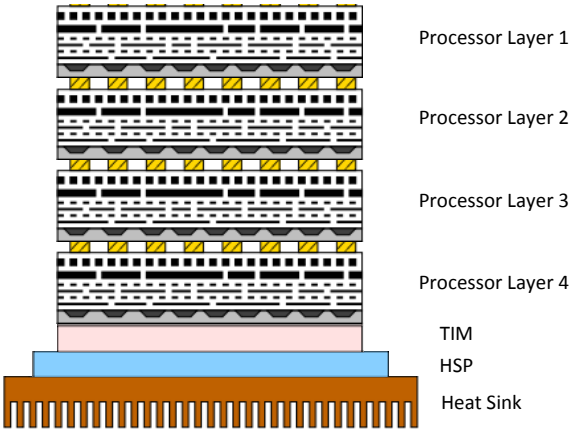

Fig. 4. 3-D thermal model (based on [11])

4 CONCLUSIONS

The associative processor (AP) is essentially a large associative memory with massively parallel processing capabilities. The PUs are uniformly spread over the area of the AP. The merit of 3-D APs is investigated in this paper, and in addition to eliminating hot spots, the AP temperature is shown to be quite low and close to uniformly distributed. A 3-D AP thus allows stacking a DRAM cube above multiple processor layers.

Associative processing architectures can potentially enable high performance 3-D stacking due to the inherent thermal advantages as compared to other massively parallel architectures such as high end GPUs.

ACKNOWLEDGMENT

We thank Kevin Skadron for his assistance with HotSpot. This research was partly funded by the Intel Collaborative Research Institute for Computational Intelligence and by Hasso-Plattner-Institut.

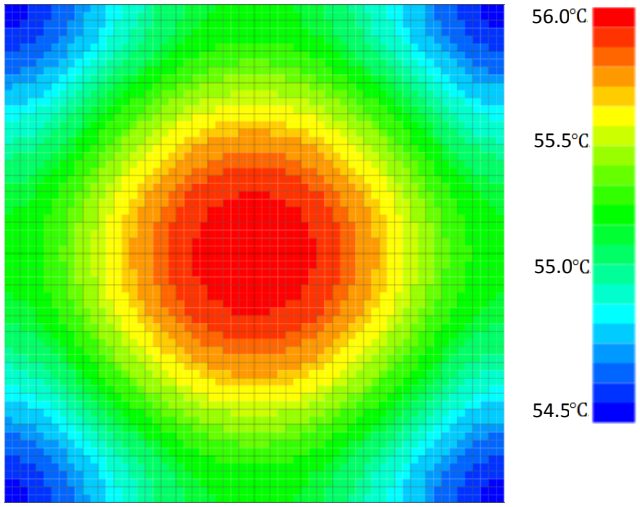

Fig. 5. A thermal map of the layer 1, 16mm^2 AP

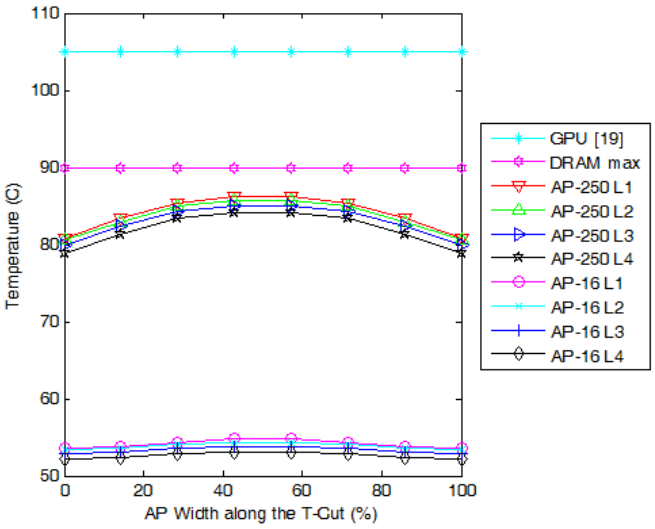

Fig. 6. Temperatures in four silicon layers along the T-Cut (cf. Fig. 3(a)) for 16 mm^2 and 250 mm^2 APs

REFERENCES

[1] A. Coskun, A. Kahng, T. Rosing, "Temperature-and cost-aware design of 3-D multiprocessor architectures", 12th Euromicro Conference on Digital System Design, Architectures, Methods and Tools, pp. 183-190, 2009

[2] Almási G. *et al.*, "Dissecting Cyclops: A detailed analysis of a multithreaded architecture", ACM SIGARCH Computer Architecture News 31.1 (2003): 26-38.

[3] Borkar S.. "Thousand Core Chips: A Technology Perspective," *Proc. ACM/IEEE 44th Design Automation Conf. (DAC)*, 2007, pp. 746-749.

[4] Burger D., T. Austin. "The SimpleScalar tool set, version 2.0", ACM SIGARCH Computer Architecture News 25.3 (1997): 13-25.

[5] C. Foster, "Content Addressable Parallel Processors," Van Nostrand Reinhold Company, NY, 1976

[6] D. Hentrich, E. Oruklu, J. Saniie. "Performance evaluation of SRAM cells in 22nm predictive CMOS technology," IEEE International Conference on Electro/Information Technology, 2009.

[7] D. Steinkraus, L. Buck, P. Simard, "Using GPUs for machine learning algorithms," IEEE ICDAR 2005.

[8] F. Black and M. Scholes, "The pricing of options and corporate liabilities," Journal of Political Economy, 81 (1973), pp. 637–654, 1973.

[9] F. Li *et al.*, "Design and management of 3-D chip multiprocessors using network-in-memory", ACM SIGARCH Computer Architecture News 34.2 (2006): 130-141.

[10] Foster C., "Content Addressable Parallel Processors", Van Nostrand Reinhold Company, NY, 1976

[11] G. Loh, "3-D-Stacked Memory Architectures for Multicore Processors," ISCA '08, pages 453–464

[12] G. Loi, et al., "A thermally-aware performance analysis of vertically integrated (3-D) processor-memory hierarchy," DAP 2006.

[13] G. Qing, X. Guo, R. Patel, E. Ipek, and E. Friedman. "AP-DIMM: Associative Computing with STT-MRAM," ISCA 2013

[14] H. Li, C. Chen, J. Wang, and C. Yeh, "An AND-type match line scheme for high-performance energy-efficient content addressable memories," IEEE Journal of Solid-State Circuits , vol. 41, no. 5, pp. 1108 – 1119, May 2006.

[15] http://www.elpida.com/en/products/index.html

[16] http://www.micron.com/products/dram

[17] http://www.samsung.com/global/business/semiconductor/product/dram

[18] I. Scherson, Kramer, Alleyne, "Bit-Parallel Arithmetic in a Massively-Parallel AP," IEEE Transactions on Computers, Vol. 41, No. 10, October 1992

[19] J. Sheaffer, K. Skadron, D. Luebke. "Studying thermal management for graphics-processor architectures," ISPASS 2005

[20] K. Banerjee et al., "A self-consistent junction temperature estimation methodology for nanometer scale ICs with implications for performance and thermal management," IEEE IEDM, 2003, pp. 887-890.

[21] K. Pagiamtzis and A. Sheikholeslami, "Content-addressable memory (CAM) circuits and architectures: a tutorial and survey," IEEE Journal of Solid-State Circuits, vol. 41, no. 3, pp. 712 – 727, March 2006.

[22] K. Puttaswamy and G. Loh, "Thermal analysis of a 3-D die-stacked high-performance microprocessor", Proceedings of the 16th ACM Great Lakes symposium on VLSI. ACM, 2006.

[23] K. Sankaranarayanan et al. "A case for thermal-aware floorplanning at the microarchitectural level," Journal of Instruction-Level Parallelism 7.1 (2005): 8-16.

[24] K. Skadron et al,. "Temperature-aware microarchitecture." APM SIGARCH Computer Architecture News. Vol. 31. No. 2. APM, 2003.

[25] L. Yavits, "Architecture and design of Associative Processor for image processing and computer vision", MSc Thesis, Technion – Israel Institute of technology, 1994, available at http://webee.technion.ac.il/publication-link/index/id/633

[26] L. Yavits, A. Morad, R. Ginosar, "Computer Architecture with Associative Processor Replacing Last Level Cache and SIMD Accelerator", IEEE Transactions on Computers, 2013

[27] L. Yavits, A. Morad, R. Ginosar, "The effect of communication and synchronization on Amdahl's law in multicore systems", Parallel Computing journal, 2013.

[28] Li H. *et al.* "An AND-type match line scheme for high-performance energy-efficient content addressable memories," IEEE Journal of Solid-State Circuits , vol. 41, no. 5, pp. 1108 – 1119, May 2006.

[29] M. Quinn, "Designing Efficient Algorithms for Parallel Computers", McGraw-Hill, 1987, page 125.

[30] R. Menon and V. Pangracious, "A Novel Methodology for Thermal Aware Silicon Area Estimation for 2D & 3-D MPSoCs", International Journal of VLSI design & Communication Systems, Vol.2, No.4, December 2011

[31] S. Choi, K. Sohn, HJ. Yoo. "A 0.7-fJ/bit/search 2.2-ns search time hybrid-type TCAM architecture", , IEEE Journal of Solid-State Circuits, 40.1 (2005): 254-260.

[32] S. Keckler, *et al.* "GPUs and the future of parallel computing", IEEE Micro, 31.5 (2011): 7-17.

[33] S. Williams, D. Patterson, L. Oliker, J. Shalf, K. Yelick, "The roofline model: A pedagogical tool for auto-tuning kernels on multicore architectures," Hot Chips 20, 2008.

[34] W. Huang, et al. "Compact thermal modeling for temperature-aware design." Proceedings of the 41st annual Design Automation Conference, APM, 2004.

[35] Yau, Fung, "AP Architecture - a Survey", APM Computing Surveys Journal (CSUR), Volume 9, Issue 1, March 1977, Pages 3 – 27