Temperature-Aware Monolithic 3D DNN Accelerators for Biomedical Applications

Prachi Shukla[1], Vasilis F. Pavlidis[2], Emre Salman[3], and Ayse K. Coskun[1]

[1]Boston University, Boston, USA - {prachis, acoskun}@bu.edu
[2]The University of Manchester, Manchester, UK - vasileios.pavlidis@manchester.ac.uk
[3]Stony Brook University, Stony Brook, USA - emre.salman@stonybrook.edu

***Abstract*—In this paper, we focus on temperature-aware Monolithic 3D (*Mono3D*) deep neural network (DNN) inference accelerators for biomedical applications. We develop an optimizer that tunes aspect ratios and footprint of the accelerator under user-defined performance and thermal constraints, and generates near-optimal configurations. Using the proposed *Mono3D* optimizer, we demonstrate up to 61% improvement in energy efficiency for biomedical applications over a performance-optimized accelerator.**

I. Introduction

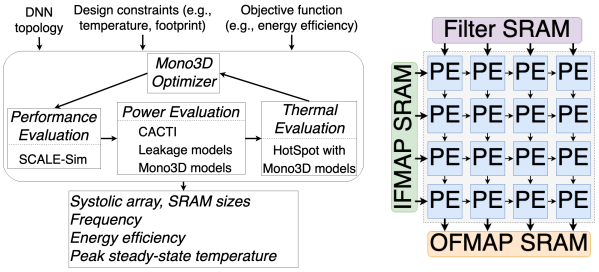
Deep neural network (DNN) inference is widely used for image segmentation and recognition in biomedical applications, e.g., improving imaging for cancer detection [1], [2]. For these applications, mobile/portable DNN accelerators are in demand to optimize for computation speed, energy efficiency, and small footprint [3]. Monolithic 3D (*Mono3D*) is an emerging 3D technology with the potential to offer these characteristics and provide improvement over 2D systems [4].

In *Mono3D* ICs, two or more thin tiers are vertically integrated in a sequential fabrication process, where nanometer-scale vias provide high-density vertical interconnects, thus leading to dense integration. Due to the thin tiers, *Mono3D* has lower vertical thermal resistance than other 3D technologies, e.g., 3D stacking [5], and results in strong inter-tier thermal coupling. Furthermore, the strong thermal coupling may lead to similar high density hot spots across tiers [6]. In addition, the absence of heat sinks and fans in mobile systems can escalate thermal concerns. Therefore, it is imperative to consider thermal awareness while designing mobile *Mono3D* systems.

To provide energy and area efficiency, while also maintaining thermal integrity in *Mono3D* systems, we utilize an existing temperature-aware optimizer to generate near-optimal mobile DNN accelerator configurations for biomedical applications. In this work, we use systolic arrays as the target DNN accelerator due to their simple architecture [7]. We investigate two DNNs (*U-Net*, *ResNet-50*) that are used for image segmentation and classification, respectively, due to their high accuracy.

II. Temperature-aware *Mono3D* Systolic Arrays

We show a temperature-aware optimization flow in Fig. 1a [8]. The inputs to the optimizer are design constraints (latency, temperature, footprint), a DNN and its topology (input/filter size, number of filters/channels, etc.), and an objective function (energy efficiency). A multi-start simulated annealer (MSA)-based optimizer iterates through performance, power, and thermal evaluation and converges to a near-optimal *Mono3D* configuration with safe chip temperature when it can no longer find better configurations. Multiple starts in MSA increase the probability of escaping local optima and converging to global optima. We show our target DNN accelerator, a systolic array, in Fig. 1b. Systolic arrays are a 2D network of processing elements (PEs) with SRAMs for input feature map (IFMAP), weights (Filter), and outputs (OFMAP). Each PE is a multiply-and-accumulate (MAC) unit with internal registers for inputs/partial sums. Inputs are read from the top and left edges and passed on to the PEs in every clock cycle (Fig. 1b).



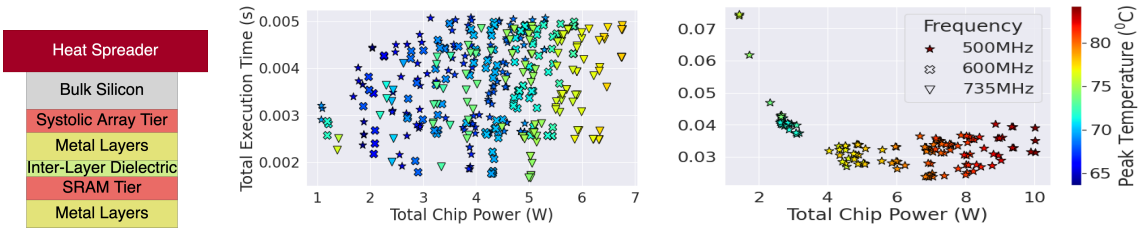
(a) *Mono3D* optimizer. (b) 4×4 systolic array.

Fig. 1: *Mono3D* optimization flow and a sample systolic array.

Several tools and models are integrated into the optimizer to evaluate *Mono3D* systolic array configurations. As shown in Fig. 1a, the optimizer starts with performance evaluation of the DNN using SCALE-Sim, a cycle-accurate stall-free DNN inference simulator for systolic arrays [9], followed by power evaluation using CACTI-6.5 [10] and *Mono3D* power models. For thermal evaluation, the optimizer uses HotSpot-6.0 to obtain steady-state temperatures [11]. There also exists a leakage-temperature loop for an accurate power/temperature estimation. The loop converges when the difference between consecutive HotSpot simulations is $< 1°C$.

We investigate a *Mono3D* configuration comprising two tiers, as shown in Fig. 2a. The logic layer, i.e., systolic array tier is closer to the heat spreader and MIVs are used for SRAM read/writes. For simplicity, we assume that the systolic array and SRAM tiers are roughly equal in size [8]. We adopt a

This paper was accepted to be presented at the Design, Automation and Test in Europe Conference (DATE) 2022 workshop on "3D Integration: Heterogeneous 3D Architectures and Sensors".



(a) *Mono3D* chipstack. (b) *ResNet-50* (c) *U-Net*

Fig. 2: Cross-sectional view of *Mono3D* chipstack (left) and performance versus power tradeoffs in *Mono3D* DNN accelerators.

representative *Mono3D* power model for interconnect power from a recent work [8]. A simplifying assumption made in this power model is that the interconnect power equals 15% of the total chip dynamic power. On top of this interconnect power, 10% interconnect power savings are applied for *Mono3D* power savings at iso-performance [8]. We also adopt a representative thermal model from a recent work [12] composed of metal layers, dielectric, etc. with the corresponding layer thicknesses and thermal resistivities.

III. Experimental Results

To demonstrate the benefits of thermal awareness in the design of DNN systolic arrays for biomedical applications, we use two DNNs: *U-Net* and *ResNet-50*. Table I shows our design space. We set a thermal threshold of 80°C and a limit on maximum performance loss of $\leq$ 10% with respect to a latency-optimized configuration. We use an example MAC unit's area, power, and frequency, and include three frequency levels in our analysis: (500, 600, 735) MHz [8]. In total, there are 6k unique configurations for each DNN, including the frequencies. We launch six starts for each frequency with five perturbations. MSA parameters are set to 1.44/0.88, 0.85 for initial/final annealing temperatures[1] and rate of cooling, respectively [8].

Table II lists the optimized configurations for inference latency, chip power, and energy-delay-area product (EDAP). We utilize EDAP to measure energy- and area-efficiency. Figures 2b and 2c show the configurations explored by the optimizer for power minimization. Absence of a frequency level depicts that the optimizer did not find a valid configuration for initialization at that frequency. As shown in the table, the optimizer converges to lowest frequency level for *U-Net* and highest frequency level for *ResNet-50*. This difference is due to the topological differences among these DNNs. *ResNet-50* downsizes the input to make a final prediction for object classification, which leads to lower systolic array utilization, lower power, and fewer thermal violations (Fig. 2b). On the other hand, *U-Net* first downsizes and then expands the input to obtain a high image resolution. Due to a larger input size in its latter layers, the array utilization is greater than in *ResNet-50*, thus leading to higher power and more thermal violations (Fig. 2c). The table also shows that due to the imposed constraint in performance loss, the optimizer converges to ≈53% larger systolic arrays for *U-Net* at 500 MHz than *ResNet-50* at 735 MHz. In comparison to latency-optimized configuration, the power- and EDAP-optimized configurations achieve 21% and 61% improvement in chip power and EDAP, respectively, while sacrificing only 9.5% in latency for *U-Net*. *ResNet-50* achieves 49% and 83% improvement in chip power and EDAP using the optimizer, while sacrificing only 7.25% in latency. We also compare these results with unoptimized points corresponding to the smallest configuration in our design space (64×68 with 352 KB SRAM) running at the lowest frequency of 500 MHz, thus characterized by low power and area. Even though these configurations have lower power (avg. 55%), the latencies are 3× (*U-Net*) and 2× (*ResNet-50*) of the fastest configurations due to fewer PEs. While the unoptimized configuration has 35% lower EDAP for *ResNet-50*, for *U-Net* this results in 50% higher EDAP due to longer latency. The above results show the importance of temperature-awareness in optimizing DNN accelerators for different objectives and DNNs. In addition, it motivates the need for systematic optimization to balance constraints and objectives in a thermally-aware manner.

Systolic array size	64×64 to 256×256
Each SRAM size	(32, 64 ... 4096) KB
Aspect ratio of the chip	0.94 to 1
Frequencies	(735, 600, 500) MHz

TABLE I: Design space for DNN accelerators.

Optimization Goal	*U-Net*	*ResNet-50*
Performance (Inference Latency)	194×192 (500 MHz) 4256 KB	186×196 (735 MHz) 4160 KB
Chip Power	162×172 (500 MHz) 3136 KB	132×138 (735 MHz) 2112 KB
System EDAP	162×172 (500 MHz) 3136 KB	134×136 (735 MHz) 2112 KB

TABLE II: Optimization results: Systolic array (operating frequency) and total SRAM (IFMAP, Filter, OFMAP).

IV. Conclusion

We demonstrate the effectiveness of including temperature-awareness in design optimization for *Mono3D* energy efficient DNN accelerators, subject to user-defined performance and thermal constraints for biomedical applications. Since *U-Net* dissipates high power and results in higher temperature, the optimizer converges to *Mono3D* configurations operating at a lower frequency for energy efficiency. For *ResNet-50*, the optimizer utilizes the thermal slack and converges to configurations operating at a higher frequency due to fewer thermal violations.

[1]Annealing temperature: Unitless MSA parameter to determine when to accept a worse solution. Rate of cooling: Decaying rate of the annealing temperature to achieve convergence.

REFERENCES

[1] O. Ronneberger *et al.*, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*. Springer, 2015, pp. 234–241.

[2] Q. A. Al-Haija and A. Adebanjo, "Breast cancer diagnosis in histopathological images using Resnet-50 convolutional neural network," in *IEMTRONICS*. IEEE, 2020, pp. 1–7.

[3] Y. Wei *et al.*, "A review of algorithm & hardware design for AI-based biomedical applications," *IEEE TBioCAS*, vol. 14, no. 2, 2020.

[4] P. Batude *et al.*, "3D sequential integration opportunities and technology optimization," in *IEEE Int. Interconnect Tech. Conf.*, 2014, pp. 373–376.

[5] X. Hu *et al.*, "Die stacking is happening," *IEEE Micro '18*, vol. 38, no. 1, pp. 22–28, 2018.

[6] P. Shukla *et al.*, "An overview of thermal challenges and opportunities for monolithic 3D ICs," 2019, pp. 439–444.

[7] H.-T. Kung, "Why systolic architectures?" *Computer*, no. 1, pp. 37–46, 1982.

[8] P. Shukla *et al.*, "Temperature-aware optimization of monolithic 3D deep neural network accelerators," in *IEEE ASP-DAC*, 2021, pp. 709–714.

[9] A. Samajdar *et al.*, "A systematic methodology for characterizing scalability of DNN accelerators using SCALE-Sim," in *IEEE ISPASS*, 2020.

[10] S. Thoziyoor *et al.*, "CACTI 6.5," *hpl.hp.com*, 2009.

[11] K. Skadron *et al.*, "Temperature-aware microarchitecture," *ACM SIGARCH Computer Architecture News*, vol. 31, no. 2, pp. 2–13, 2003.

[12] C. Yan and E. Salman, "Mono3D: Open source cell library for monolithic 3-D integrated circuits," *IEEE TCAS I*, vol. 65, no. 3, 2018.