# Temperature-Aware Sizing of Multi-Chip Module Accelerators for Multi-DNN Workloads

Prachi Shukla*†, Derrick Aguren†, Tom Burd†, Ayse K. Coskun*, John Kalamatianos†
* Boston University - (prachis, acoskun)@bu.edu
† Advanced Micro Devices - (prachi.shukla, derrick.aguren, tom.burd, john.kalamatianos)@amd.com

***Abstract*—This paper demonstrates the need for temperature awareness in sizing accelerators to target multi-DNN workloads. To that end, we build TESA, a TEmperature-aware methodology that Sizes and places Accelerators to balance both the cost and power of a multi-chip module (MCM), including DRAM power for multi-deep neural network workloads. TESA tunes the accelerator chiplet size and inter-chiplet spacing to generate a temperature-aware MCM layout, subject to user-defined latency, area, power, and thermal constraints. Using TESA for both 2D and 3D systolic array-based chiplets, we demonstrate up to 44% MCM cost savings and 63% DRAM power savings, respectively, over a temperature-unaware baseline at iso-frequency and iso-interposer area. We also demonstrate a need for TESA to obtain feasible MCM configurations for multi-DNN workloads such as augmented/virtual reality (AR/VR).**

***Index Terms*—3D stacking, thermal awareness, multi-DNN workloads, systolic arrays, multi-chip module**

## I. Introduction

Deep neural networks (DNNs) are extensively used for inference in several emerging edge applications, including autonomous vehicles, augmented/virtual reality (AR/VR), etc. In these applications, multiple independent DNNs execute independent subtasks, such as speech or object recognition, to complete one large task under latency constraints [1]. Thus, these DNNs do not require inter-DNN communication to complete their tasks. Also, topological differences among DNNs [1] impact latencies and accelerator utilization, leading to varying performance, power, and thermal profiles.

To meet the latency constraints of these multi-DNN workloads, corresponding multi-accelerator systems with individual DNNs executing in parallel on distinct hardware are desirable. Since large, monolithic, multi-accelerator systems are expensive, multi-chip module (MCM)-based solutions are an alternative choice to improve yield and reduce fabrication cost [2]. Such chiplet-centric design opens additional degrees of freedom for co-optimizing performance, power, and thermals–namely, chiplet size[1] (#PEs and SRAMs), quantity, and placement, while holding total MCM area[2] fixed.

Fig. 1 shows example scenarios that motivate the need for thermal awareness in chiplet design. Usually, chiplets are placed close to one another on an interposer to shorten communication links for better performance and to improve area utilization for reducing MCM cost[3] [3] (Fig. 1a). However, a dense layout may cause high thermal coupling between neighboring chiplets and lead to high temperatures, potentially thermally constraining MCM performance. Hence, spreading the chiplets out can alleviate the hot spots. Under area constraints, spreading may only be possible by shrinking the chiplets by putting fewer processing elements (PEs) in each chiplet, thus allowing more whitespace (Fig. 1b). However, smaller chiplets with fewer PEs may hurt latency and throughput. On the other hand, increasing the chiplet size by adding more PEs will lead to fewer chiplets (Fig. 1c) and may result in higher power consumption and temperature due to a larger number of active PEs. Thus, there exist tradeoffs between various design decisions. Hence, finding an MCM configuration that satisfies all constraints is often challenging (Fig. 1d), which is the focus of this work.

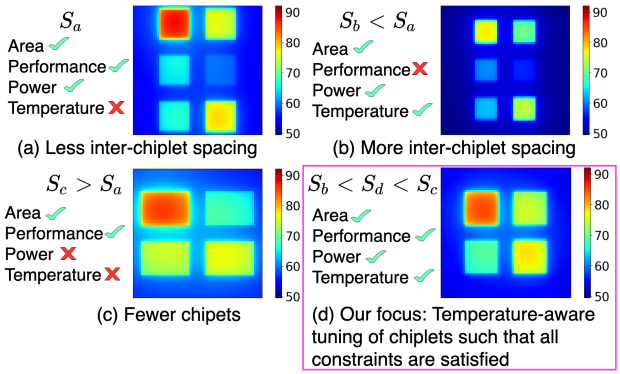
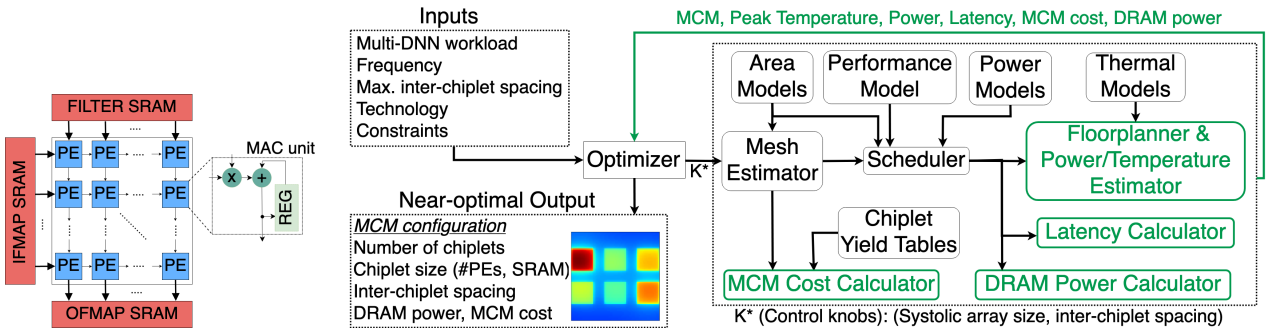



Fig. 1: An example showing several scenarios for tuning chiplet design of an MCM. $S_i$ denotes size of a chiplet in each scenario. (a) violates thermal constraint, (b) violates performance constraint due to smaller chiplets with fewer PEs, (c) violates power and thermal constraints due to more active PEs, (d) our work's focus: Temperature-aware tuning of chiplet size and placement to satisfy all of the constraints.

Prior works on 2.5D floorplanning optimize for temperature, but assume fixed chiplet architecture and number of chiplets [3], [4]. They may also be insufficient to address the full complexity of the design space: they do not treat chiplet size and quantity as variables, and do not consider how DRAM power, which substantially affects system power [5], is influenced by chiplet configuration. Nor do these works tune domain-specific (DS) chiplet architectures to meet a workloads's performance, area, power, and thermal constraints.

To address the research gap, we build TESA to demonstrate that temperature-aware chiplet sizing is critical in designing DS MCMs. We investigate both 2D and two-tier SRAM-stacked 3D chiplets motivated by AMD's recent venture into AMD 3D V-Cache™ [6]. We apply TESA to systolic arrays (Fig. 2a) as a test case, noting that the same method can be readily extended to other accelerator engines. Our contributions are as

[1]Sizing a chiplet implies selecting the chiplet architecture (i.e., #PEs and SRAM capacity). Hence, sizing impacts chiplet performance, power, and area.
[2]MCM area is the same as the interposer area in this paper.
[3]MCM cost is the fabrication cost of the system. It is a function of chiplet area and yield, microbumping cost, and the interposer.




(a) A sample systolic array.

(b) Flow diagram for TESA.

Fig. 2: (a) A systolic array with PEs and three SRAMs. Each PE is a MAC unit with internal registers. (b) TESA's flow diagram.

follows: **(1)** we demonstrate a need for thermally-aware chiplet sizing for multi-DNN workloads. To that end, we introduce TESA, in which we develop an optimizer that generates MCMs while satisfying user-defined temperature, area, performance (latency), and power constraints. The optimizer considers chiplet size (#PEs and SRAM capacity), inter-chiplet spacing (ICS), and 3D-stacking as parameters to generate an MCM configuration for the given workload. **(2)** We build detailed models of performance, power, area, temperature, MCM cost, and DRAM power into TESA. These models enable it to evaluate the impact of chiplet size, chiplet quantity, ICS, and operating frequency on multi-DNN inference on 2D/3D MCMs.

For systolic array-based MCMs, TESA demonstrates up to 44% MCM cost and 63% DRAM power savings over baselines that maximize parallelism but do not consider temperature. TESA also generates 3D MCMs with 39% higher operations per second ($OPS$).

## II. Related Work

**Accelerators for multi-DNN workloads.** Previous works on accelerator design for multi-DNN workloads do not consider thermal awareness. Nor do they consider the strict packaging constraints, in terms of area, power, and thermal limits, found in several mobile/edge platforms such as drones or AR/VR. E.g., a multi-accelerator system with different dataflows was proposed for low latency and energy efficiency [1]. DNNs and hardware architecture were co-optimized to design multi-DNN systems for high accuracy and efficiency [7]. Other works have reduced inference latency using multi-FPGA systems [8] or CPU-GPU hybrid architecture [9].

**Thermal awareness in 2.5D/3D.** 2.5D/3D offers 'More than Moore' but can face thermal challenges. Prior works on temperature-aware 2.5D floorplanning target general-purpose (GP) systems and workloads. E.g., temperature and network latency have been co-optimized in a fixed MCM [3]. A temperature-aware floorplanning tool was proposed but lacked performance and leakage models [4]. However, these works are insufficient for designing MCMs for multi-DNN workloads with specific design and performance constraints. Nor do they consider temperature-aware tuning of chiplet size to meet package/workload constraints. Another body of work investigates only the die-level thermal behavior of 3D systolic arrays [10], [11], [12].

Collectively, these works have the following limitations: (i) none of them investigates the effect of chiplet sizing (#PEs, SRAM capacity) and placement on thermal integrity and feasibility of multi-accelerator systems; (ii) nor do they account for the effect of chiplet size on DRAM power and MCM cost; and (iii) they do not consider multi-DNN workloads or design a system for such workloads, while also considering the power, area, latency, and thermal constraints together. We present a novel method, TESA, which addresses all of the above.

## III. TESA

TESA utilizes temperature for designing MCMs targeting multi-DNN workloads. The typical approach is to put as many chiplets as possible in a given MCM area, assuming a chiplet size (#PEs, SRAM capacity), to enable high parallelism. However, to reduce thermal coupling between chiplets in case of a thermal violation, ICS may be increased by making the chiplets smaller, by decreasing the systolic array and SRAM sizes. While smaller chiplets may reduce MCM cost, frequent off-chip DRAM accesses can occur due to smaller SRAMs. Alternatively, larger chiplets may access the DRAM less frequently due to better data reuse in larger SRAMs but have a higher cost. TESA considers this tradeoff. We use a representative MCM cost model that jointly accounts for chiplets, interposer, and microbump bonding cost, assuming known good dies [3]. We use systolic arrays as our test case (Fig. 2a), which is a 2D PE array with SRAMs for inputs (IFMAP), filter weights (FILTER), and outputs (OFMAP). In each cycle, inputs and weights are read from the SRAMs to compute partial sums, which are passed to neighboring PEs for accumulation and written to the OFMAP SRAM [13]. The rest of the section starts with TESA's overview, followed by its models, our workload policy, and optimizer.

### A. TESA *Overview*

Fig. 2b shows TESA's flow diagram. The inputs to TESA are (i) multi-DNN workload (layer-wise description of each DNN with input size, #weights, etc.), (ii) frequency, (iii) maximum allowed ICS, (iv) integration technology (2D or 3D), and (v) constraints on latency, power, interposer area, and temperature. TESA then uses chiplet size and ICS as control knobs to output an MCM to balance cost and DRAM power. To efficiently traverse and evaluate the design space, we integrate a multi-start simulated annealing (MSA)-based optimizer into TESA. For simplicity, to focus on the methodology, the optimizer fills the interposer area uniformly with chiplets,

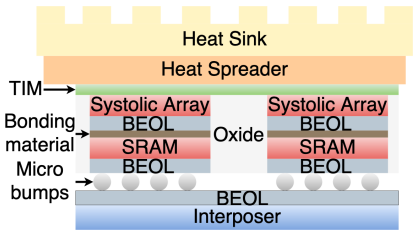



Fig. 3: Cross-sectional view of 3D chiplets on an interposer with the SRAMs stacked underneath systolic arrays in a F2B manner. TSVs provide vertical interconnection between the two tiers. On the other hand, in 2D chiplets, a systolic array and its SRAMs are placed adjacent to one another.

which results in a dense mesh-like layout. The optimizer first introduces random perturbations to chiplet size and ICS. A mesh estimator then generates a mesh (rows×columns) for the given chiplet size and ICS. We limit the number of chiplets to the number of DNNs in the workload to avoid over-provisioning the accelerator. A scheduler then assigns DNNs to chiplets using a latency, power, and power-density aware policy. Using our power models, mesh, and schedule, a floorplanner generates an MCM. TESA also has leakage and thermal models to estimate total power and temperature. In parallel, TESA calculates DRAM power, MCM cost, and latency. Finally, the optimizer uses the generated outputs, namely peak temperature, total power, latency, cost, and power, to evaluate the MCM. The optimizer iterates through this flow and converges to a near-optimal MCM.

**Architecture and Multi-DNN Workload.** We investigate MCMs of systolic arrays with SRAMs on a silicon interposer. Fig. 3 shows a cross-sectional view of a 3D chiplet. Each chiplet is assumed to have independent DRAM channels. The number of channels assigned to a chiplet is determined by its bandwidth requirements. We also use a user-defined ICS constraint. We investigate an AR/VR workload of six DNNs to perform the following independent tasks: handpose detection (*HandposeNet*), image segmentation (*U-Net*), object detection (*MobileNet*), object recognition (*ResNet-50*), depth estimation (*DNL*), and speech recognition (*Transformer*). The first five are taken from a representative AR/VR workload [1]. We add *Transformer* for speech recognition [14]. We assume ICS does not affect the overall latency because: (i) there is no need for inter-DNN communication since each DNN performs an independent subtask [1], and (ii) the chiplets are placed along the edges and have dedicated DRAM channels. Thus, ICS does not significantly impact DRAM latency.

### B. Models

**Performance models.** We use a DNN simulator, SCALE-Sim, that models stall-free inference on systolic arrays with double-buffered SRAMs [13]. SCALE-Sim takes the DNN topology, systolic array size, SRAM size, and dataflow as inputs. It then runs cycle-accurate stall-free inference on 8-bit integer data. It outputs a performance summary, such as execution cycles, systolic array utilization, and average/peak DRAM and SRAM bandwidths, assuming a batch size of 1. Both 8-bit data and batch size=1 hold for AR/VR workloads [1]. The performance model is the same for 2D and 3D MCMs because our analysis is at iso-frequency.

TABLE I: Notations used in Equations (1) to (5).

| Notation | Meaning |
|---|---|
| $\text{Chip}_i, \text{DNN}_j$ | Chiplet $i$, DNN $j$ |
| $\text{DP}_{i,\ j}$ | $\text{Chip}_i$'s dynamic power when executing $\text{DNN}_j$ |
| $\text{SaDP}_{i,\ j}$ | $\text{Chip}_i$'s systolic array dynamic power for $\text{DNN}_j$ |
| $\text{SrDP}_{i,\ j}$ | $\text{Chip}_i$'s total SRAM dynamic power for $\text{DNN}_j$ |
| $\text{Util}_{i,\ j}$ | Average systolic array utilization (%) in $\text{Chip}_i$ by $\text{DNN}_j$ |
| freq | Operating frequency of systolic arrays |
| $\text{DP}_{MAC,freq}$ | Dynamic power of a MAC unit at freq |
| num_$PEs_i$ | Number of PEs in $\text{Chip}_i$ |
| $\text{L}_j$ | No. of CNN and FC layers in $\text{DNN}_j$ |
| $\text{Util}_{i,\ j,\ k}$ | $\text{Chip}_i$'s systolic array utilization by $\text{DNN}_j$'s $\text{k}^{th}$ layer |
| $\text{CC}_{i,\ j,\ k}$ | Compute cycles of $\text{DNN}_j$'s $\text{k}^{th}$ layer on $\text{Chip}_i$ |
| $\text{Sr}_m$ | IFMAP, FMAP, OFMAP for $m$=1, 2, 3, respectively |
| $\text{SrBw}_{avg,m}$ | Average $\text{Sr}_m$ bytes accessed per cycle |
| $\text{DP}_{Sr_m,byte}$ | Dynamic power per $\text{Sr}_m$ byte access from CACTI-7.0 [15] |
| $\text{TSV}_{power,bit}$ | TSV's dynamic power/bit (1 $\mu$W at 400 MHz) [16] |
| $\text{TsvDP}_{i,\ j}$ | TSV's dynamic power for $\text{Chip}_i$ running $\text{DNN}_j$ |

**Power models.** We calculate the dynamic power ($\text{DP}_{i,j}$) for each chiplet $\text{Chip}_i$ (i.e., systolic array and SRAMs) running a DNN $\text{DNN}_j$ using the frequency and SCALE-Sim outputs:

$$DP_{i,j} = SaDP_{i,j} + SrDP_{i,j}[10], \tag{1}$$

$$SaDP_{i,j} = Util_{i,j} \cdot DP_{MAC,freq} \cdot num\_PEs_i[10], \tag{2}$$

$$Util_{i,\ j} = \frac{\sum_{k=1}^{L_j} Util_{i,\ j,\ k} \cdot CC_{i,\ j,\ k}}{\sum_{k=1}^{L_j} CC_{i,\ j,\ k}}, \tag{3}$$

$$SrDP_{i,\ j} = \sum_{m=1}^{3} SrBw_{avg,m} \cdot DP_{Sr_m,byte}, \tag{4}$$

where Table I lists all the notations. We use CACTI-7.0 for SRAM leakage and a representative exponential model for systolic array's leakage [10]. Different from 2D, in 3D MCMs, we add TSV power (TsvDP) using average SRAM bandwidth ($\text{SrBwS}_{avg}$) for each SRAM (IFMAP, FILTER, and OFMAP), as shown in Eq. (5), to the SRAM tier's back end of line (BEOL). Using a representative energy/bit value [16], we calculate power/bit ($\text{TSV}_{power,bit}$) at frequency *freq*.

$$TsvDP_{i,\ j} = \sum_{m=1}^{3} SrBw_{avg,m} \cdot TSV_{power,bit} \cdot 8 \tag{5}$$

We use Micron's DRAM power model for DDR4 SDRAM, which includes the refresh power, standby power, I/O power, etc. 2D and 3D chiplets have the same model. We assume each chiplet has independent channels. If a chiplet runs multiple DNNs sequentially, the highest number of channels across those DNNs is assigned to it. Microbump power is ignored [17].

**Area model.** We make two simplifying assumptions: (i) the area ratio of a systolic array to its three SRAMs is ≈1 [10], and (ii) each of the three SRAMs is of the same size [12]. We use CACTI-7.0 for SRAM area estimates and a representative area estimate for MAC units [10]. In 3D, the SRAM tier has an additional TSV area overhead. The peak SRAM bandwidth determines the number of TSVs in a 3D chiplet. The footprint of a 3D chiplet is $max$(SRAM tier area, systolic array tier area). We use aggressive TSV dimensions with diameter and keep-out-zone as 2 $\mu m$ [18].

**Thermal model.** We use HotSpot-6.0 for steady state thermal simulations [19]. Fig. 3 shows a cross-sectional view of our model. The material properties are from prior work [20]. We

use the default HotSpot ambient temperature (45°C) and set convection resistance to 0.4 K/W to represent limited cooling in edge/mobile devices [19]. The TSVs pass through the SRAM tier in 3D chiplets, so we use copper and silicon joint resistivity to estimate the SRAM tier's thermal resistance based on TSV area occupancy. If the number of chiplets is fewer than the number of DNNs in the workload, TESA performs steady state analysis for each set of DNNs that can execute simultaneously on different chiplets at any given time until convergence, and then reports the maximum temperature.

### C. Workload scheduling policy

We build a deterministic, latency-, power-, and power-density-aware static scheduling policy that assumes non-preemptive DNN execution, similar to a prior work that considers non-preemptive scheduling for energy savings in low-power devices [21]. Thus, a DNN finishes executing before another DNN begins execution on the same chiplet. The DNNs are first assigned to the chiplets in the corner, followed by outer rows/columns, and then to the center to avoid hot spots. If the number of chiplets is fewer, remaining DNNs are scheduled greedily to idle chiplets. Both execution cycles and array utilization are obtained from SCALE-Sim.

### D. Optimizer

We construct an MSA-based optimizer (see Fig. 4) that converges to a near-optimal MCM to minimize an objective function (Eq. (6)), Obj, while satisfying the user-defined

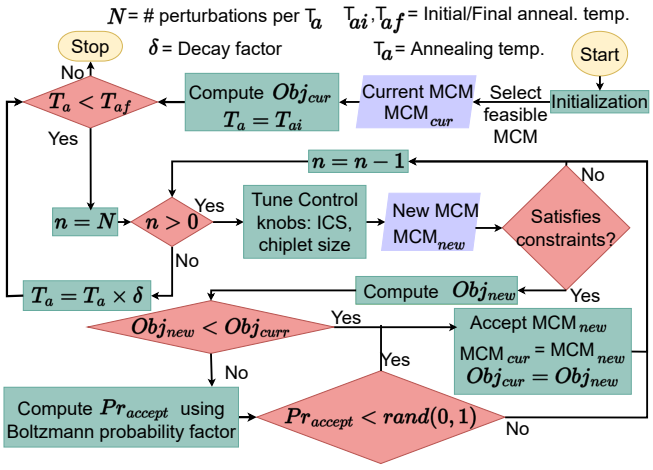



Fig. 4: Flow diagram of an annealer in TESA. Annealers are defined using annealing tempertures[a] and decay rate. Each annealer starts with initialization with a feasible MCM[b] and computes the objective function. The optimizer performs $N$ perturbations to traverse the design space and finds a feasible MCM that minimizes the objective function and finally converges[c]. In each perturbation, it either tunes the chiplet size or ICS to generates a new MCM. If the new MCM does not satisfy the constraints, the optimizer rejects it and moves to the next iteration. Otherwise, if the $\text{MCM}_{new}$ is a better configuration ($Obj_{new} < Obj_{cur}$) the optimizer accepts it, updates $Obj_{cur}$, and moves to the next iteration. However, if $Obj_{new} > Obj_{curr}$, it generates a probability of acceptance ($Pr_{accept}$), and accepts $\text{MCM}_{new}$ if $Pr_{accept}$ is smaller than a random number uniformly drawn between 0 and 1. $T_a$ reduces by $\delta$ after $N$ perturbations and the optimizer converges when $T_a > T_{af}$.

[a]Decides whether to accept a worse configuration to escape a local minima.
[b]Feasible MCMs are those that satisfy all of the user-defined constraints.
[c]The optimizer converges when it can no longer find better solutions that minimizes an objective function.

TABLE II: Design space and user-defined constraints.

| | |
|---|---|
| Systolic array size | 16x16 to 256x256, Aspect ratio =1 |
| SRAM Size | 8,16, ... 4096 $KB$ |
| ICS | Min = 0, Max = 1 $mm$ |
| Frequencies | 400 MHz, 500 MHz |
| Interposer area constraint | 8 $mm$ × 8 $mm$ |
| Thermal budgets | 75°$C$, 85°C |
| Latency constraints | 15 $fps$, 30 $fps$ [22] |
| Power budget | 15 $W$ [23] |

performance, power, area, and temperature constraints. Multiple starts execute in parallel and increase the probability of reaching the global optima.

$$Obj = Min : \alpha \cdot MCMcost_{normalized} + \beta \cdot DRAMpower_{normalized}, \quad (6)$$

where $\alpha$, $\beta$ are user-defined weights. $MCMcost_{normalized}$ and $DRAMpower_{normalized}$ are normalized MCM cost and DRAM power, respectively.

## IV. Experimental Results

This section demonstrates the importance of temperature in chiplet sizing in edge/mobile DNN systolic arrays. We discuss the design space and optimizer's correctness and compare TESA to other works and baselines. We also apply TESA to compare 2D and 3D MCMs.

### A. Experimental Setup

**Design Space.** Table II lists a representative edge/mobile design space with constraints. There are 121 systolic arrays (16×16, 18×18 ...) that span commonly used sizes [24], [13]. To simplify the design space, we discretize ICS using a small step size of 50 $\mu m$ and obtain 21 discrete options. Since we are analyzing a six-DNN workload, there are 14 meshes (1×1, 1×2,...6×1). In total, there are 35.6k unique MCMs. We set representative constraints on latency, power, and temperature. We use an arbitrary interposer area constraint to represent area constraints in edge devices. We explore two frequencies to demonstrate the impact of temperature on chiplet sizing [25]. Note that our work aims to show the importance of temperature in designing MCMs for multi-DNN workloads. We do so by applying TESA to a representative design space. Alternatively, we can apply TESA to a different and larger design space per workload/chip design requirements. We obtain representative dynamic power, leakage, and area estimates for a 22 nm MAC [10] and use CACTI-7.0 for 22 nm SRAM estimates. We enable *detailed_3D* in HotSpot for heterogeneous simulations with 125 $\mu$m grids.

**Optimizer's runtime evaluation and correctness.** SCALE-Sim simulation time varies between tens of minutes (e.g., *ResNet-50* on 256×256 array) to 12 hours (e.g., *U-Net* on 16×16 array) due to varying DNN topologies and chiplet sizes. HotSpot steady state simulation takes approximately 6 s and 16 s for 2D and 3D MCMs, respectively. Temperature-leakage convergence takes up to 3 and 6 HotSpot iterations in 2D and 3D MCM layouts, respectively. An exhaustive evaluation can take multiple days for this DNN workload and thus, there is a need for an optimizer.

To validate our optimizer's correctness, we first exhaustively evaluate a small design space with 5k MCMs and determine the globally optimum 2D and 3D MCMS for $\alpha$=$\beta$=1 to

TABLE III: Comparison of TESA to prior works at 500 MHz. The design space and other constraints used here are listed in Table II.

| Method | Adoption of original method | Adoption of method with performance & power constraints |
|---|---|---|
| W1 [4]<br>- Fixed chiplet power<br>- *Obj*: Minimize T<br>- Constraints: None<br>- No perf. model | 1×6 grid of 16×16 array with 24 KB SRAM chiplets,<br>ICS = 800 $\mu$m<br>Infeasible MCM: Performance constraint violation<br>(latency 36× longer than 30 fps)<br>**Chiplet resizing is needed to guarantee the desired perf.** | 2×3 grid of 132×132 array with 1,536 KB SRAM,<br>ICS = 1 mm<br>Infeasible MCM: Thermal constraint violation at 75°C<br>(Peak temp. = 81°C)<br>**Chiplet resizing is needed to meet the 75°C constraint** |
| W2 [3]<br>- *Obj*: Minimize T,<br>MCM cost, latency<br>- Constraints: None | 3×2 grid of 56×56 array w/ 192 KB SRAM chiplets,<br>ICS=900 $\mu$m<br>Infeasible MCM: Performance constraint violation<br>(latency 4× longer than 30 fps)<br>**Latency minimization does not guarantee desired perf.** | 2×3 grid of 130×130 array w/ 1,536 KB SRAM chiplets, ICS=1 mm<br>Infeasible MCM: Thermal constraint violation at 85 and 75°C<br>(Peak temp. = 88°C)<br>**Power constraint does not guarantee safe temperatures due to increased vertical and lateral thermal coupling** |
| TESA | Solution does not exist at 75°C; This is an important result for system designers to make remedial decisions (e.g., reduce frequency) | |

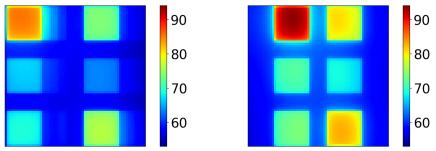
balance cost and DRAM power. The validation design space contains 64×64 to 128×128 arrays and a coarse 200 $\mu$m ICS step size. We use these constraints - 15 W, 15 fps, and 85°C, and both frequencies. The optimizer has three starts with $\delta$=0.89, 0.87, and 0.85, respectively. Each annealer has following properties: $T_{ai}$ = 0.5, $T_a$ = 19, $N$=10. Thus, the final probabilities of accepting worse solutions are very low, e.g., $2\times10^{-6}$ for $\delta$=0.85. The optimizer explores <15% of this design space before convergence with a 100% agreement with the global optima. Thus, we ensure the optimizer shows close agreement with the global optima. In Sec. IV-B, we use these MSA parameters in our original design space (Table II).

### B. Results

We evaluate the major benefits of TESA against two prior 2.5D floorplanning methodologies-$W1$ [4], and $W2$ [3], and temperature-unaware theoretical baselines.

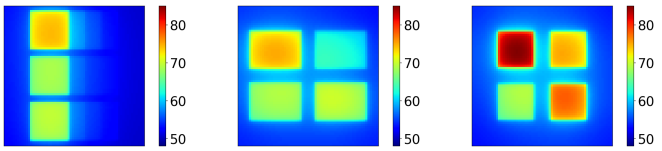
*1) Comparison with $W1$ and $W2$:* There are three major differences between TESA and these works. **First**, they target GP applications and GP chiplet architecture, thus, are insufficient to target *DS applications*, e.g., AR/VR, because: (i) DS applications have strict performance constraints. Furthermore, mobile platforms have additional power, area, and thermal constraints, which they do not sufficiently consider. (ii) GP hardware may be sub-optimal for DNNs than specialized accelerators [5], [24]. In contrast, TESA generates *DS MCMs* for multi-DNN workloads. It can re-evaluate the design space and output a different MCM for a different multi-DNN workload. **Second**, unlike TESA, they do not investigate *chiplet architecture* and instead fix the MCM compute capability, i.e., chiplet size, the number of chiplets, and/or chiplet power. **Third**, unlike these works, TESA models *3D chiplets* with detailed models. Also, leakage is vital in temperature estimation, especially in 3D, because thermal coupling can cause a significant rise in leakage, resulting in a thermal runaway situation [26]. The above works either ignore or [4] or under-estimate leakage with a linear model [3], which can result in infeasible/sub-optimal MCMs.

Table III compares these works for 3D MCMs. We modify TESA to use the methods presented in W1 and W2. *W1 and W2 do not perform temperature-aware tuning of chiplet size and quantity.* So, for a fair comparison, we fix the number of chiplets to six to enable parallel DNN execution for low latency. The table shows that these works lead to thermally infeasible MCMs and cannot guarantee the desired performance for DS applications. Thus, there is a need for thermal awareness, along with appropriate performance and power models.




(a) 2D baseline at 500 MHz.(b) 3D baseline at 500 MHz.

Fig. 5: SC1 MCMs that maximize parallelism without thermal awareness. Each chiplet is a 180×180 array w/ 1,536 KB SRAM. All MCMs, including at 400 MHz, exceed 75°C. (b) violates power budget due to ignoring leakage.




(a) 400 MHz, 30 fps, 75°C.(b) 400 MHz, 30 fps, 75°C.(c) 500 MHz, 15 fps, 85°C.

Fig. 6: Thermal maps: A subset of TESA outputs at (400, 500) MHz. (a) is a 2D MCM output, (b)-(c) are 3D MCM outputs.

*2) Comparison to temperature-unaware baselines:* We build two hypothetical baselines: (i) $SC1$ for temperature-unaware MCMs that maximize parallelism, i.e., each DNN runs simultaneously on a dedicated chiplet to avoid latency constraint violation. Even though it is temperature-unaware, it uses the maximum ICS (1 mm) to reduce lateral thermal coupling effects that escalate temperature (see Fig. 5). (ii) $SC2$ for chiplet sizing without temperature. We disable the thermal and leakage models in the baselines, and the power constraint is applied only to dynamic power. Table IV lists the SC2 outputs. Under a strict 75°C budget, it fails to produce feasible 2D MCMs at 500 MHz. For 3D MCMs, all outputs except one lead to thermal runaway conditions due to the lack of a leakage model and temperature evaluation. Thus, temperature-aware chiplet sizing is critical for both 2D and 3D MCMs.

In contrast to SC1 and SC2, TESA always outputs feasible MCMs due to thermal awareness. Table V lists TESA's outputs, while Fig. 6 shows thermal maps for a subset of these outputs for $\alpha$=$\beta$=1 in Eq. (6) to balance both cost and DRAM power. TESA finds appropriate MCMs for different latency constraints and captures chiplet sizing trends. E.g., at iso-frequency and iso-technology, the array size at 75°C is smaller or equal to those at 85°C due to lower power dissipation. However, at 500 MHz in 2D, TESA selects a larger 240×240 array at 75°C. Despite 16% more power than 200×200 array at 500 MHz, its area is 20% larger, which results in lower power density and 2.3°C lower temperature.

Compared to SC2, TESA ensures the thermal integrity of MCMs by generating smaller chiplets. However, while the

TABLE IV: SC2's 2D/3D MCMs: Chiplet sizing without thermal awareness.

| Chiplet Architecture and Tech. | Grid size, ICS | Frequency, performance constraint | Peak Junction Temp. |
|---|---|---|---|
| 200×200 array 3,072 KB SRAM (2D) | $2\times1, 850\ \mu m$ | 400 MHz, 15 fps | 72.01°C |
| | $3\times1, 650\ \mu m$ | 400 MHz, 30 fps | 74.46°C |
| | $2\times1, 700\ \mu m$ | 500 MHz, 15 fps | 77.53°C |
| | $2\times1, 900\ \mu m$ | 500 MHz, 30 fps | 77.34°C |
| 216×216 array 3,072 KB SRAM (3D) | $2\times2, 550\ \mu m$ | 400 MHz, 15 fps | Thermal runaway |
| | $2\times2, 900\ \mu m$ | 400 MHz, 30 fps | 80.03°C |
| | $2\times2, 600\ \mu m$ | 500 MHz, 15 fps | Thermal runaway |
| | | 500 MHz, 30 fps | Thermal runaway |

TABLE V: TESA's outputs: 2D/3D MCMs at (400, 500) MHz and constraints.

| Architecture and Tech. | Grid size, ICS | Frequency, constraints | Peak Temp. |
|---|---|---|---|
| 200×200 array 3,072 KB SRAM (2D) | $2\times1, 700\ \mu m$ | 400 MHz, 15 fps, 75°C | 72.11°C |
| | $2\times1, 750\ \mu m$ | 400 MHz, 15 fps, 85°C | 72.08°C |
| | $3\times1, 400\ \mu m$ | 400 MHz, 30 fps, 75°C | 74.38°C |
| | $3\times1, 250\ \mu m$ | 400 MHz, 30 fps, 85°C | 74.47°C |
| | $2\times1, 700\ \mu m$ | 500 MHz, 15 fps, 85°C | 77.53°C |
| | $2\times1, 950\ \mu m$ | 500 MHz, 30 fps, 85°C | 77.30°C |
| 240×240 array 3,072 KB SRAM (2D) | $2\times1, 950\ \mu m$ | 500 MHz, 15 fps, 75°C | 74.99°C |
| | $2\times1, 950\ \mu m$ | 500 MHz, 30 fps, 75°C | 74.99°C |
| 196×196 array 3,072 KB SRAM (3D) | 2×2, 1 mm | 400 MHz, 15 fps, 75°C | 74.64°C |
| | $2\times2, 800\ \mu m$ | 400 MHz, 30 fps, 75°C | 74.99°C |
| 216×216 array 3,072 KB SRAM (3D) | $2\times2, 700\ \mu m$ | 400 MHz, 15 fps, 85°C | 82.30°C |
| | $2\times2, 800\ \mu m$ | 400 MHz, 30 fps, 85°C | 80.88°C |
| 96×96 array 768 KB SRAM (3D) | $3\times2, 950\ \mu m$ | 500 MHz, 15 fps, 75°C | 73.66°C |
| 186×186 array 1,536 KB SRAM (3D) | $2\times2, 950\ \mu m$ | 500 MHz, 15 fps, 85°C<br>500 MHz, 30 fps, 85°C | 84.88°C |

MCM cost improves by 17%, DRAM power increases by 37.8%. Compared to $SC1$, TESA achieves a 63% and 44% reduction in DRAM power and cost due to chiplet resizing.

*3) Comparison of 2D and 3D MCMs:* We compare 2D and 3D MCMs outputs of TESA (Table V), averaged over both frequencies. We also compare $OPS$, a standard DNN inference metric, where each MAC operation equals two operations. 3D MCMs can provide up to 39% better $OPS$, on average, while sacrificing 61% in MCM cost and 66% in DRAM power at the relaxed 85°C constraint. Due to footprint savings in 3D chiplets, TESA places a larger number of chiplets on the interposer, thus, resulting in better OPS but higher DRAM power due to simultaneous DNN execution. However, 3D MCMs have higher costs resulting from the additional stacking and microbumping cost due to more chiplets. Interestingly, $OPS$ improvement is more significant at 85°C than at 75°C because TESA utilizes the thermal headroom and resizes the chiplets to larger systolic arrays. This demonstrates the advantage of thermally tuning the chiplet size and spacing to utilize the available thermal headroom.

## V. Conclusions and Discussion

We demonstrate that temperature awareness is critical for designing MCMs for multi-DNN workloads. We build TESA, a novel method to include temperature into chiplet sizing to balance MCM cost and DRAM power. We show that existing works are insufficient in determining feasible MCMs for multi-DNN workloads. Without TESA, thermal trends and phenomena in 2D and 3D MCMs may not be easily captured. TESA can help chip designers identify thermally infeasible solutions and take remedial decisions, e.g., reducing frequency. Compared to temperature-unaware baselines, TESA achieves up to 44% cost and 63% DRAM power savings. Compared to 2D MCMs, 3D MCMs achieve up to 39% $OPS$ improvement. As future work, we plan to integrate thermal awareness into other works [1], [2] for a complete system design that includes a network-on-package, continuous ICS values, or more frequencies [25], etc.

## References


[1] H. Kwon *et al.*, "Heterogeneous dataflow accelerators for multi-DNN workloads," in *IEEE HPCA*, 2021, pp. 71–83.

[2] Y. S. Shao *et al.*, "Simba: Scaling deep-learning inference with multi-chip-module-based architecture," in *IEEE/ACM MICRO*, 2019, pp. 14–27.

[3] A. Coskun *et al.*, "Cross-layer co-optimization of network design and chiplet placement in 2.5D systems," *IEEE TCAD*, vol. 39, no. 12, pp. 5183–5196, 2020.

[4] Y. Ma *et al.*, "TAP-2.5D: A thermally-aware chiplet placement methodology for 2.5D systems," in *IEEE DATE*, 2021, pp. 1246–1251.

[5] V. Sze *et al.*, "Efficient processing of deep neural networks: A tutorial and survey," *Proc. of the IEEE*, vol. 105, no. 12, pp. 2295–2329, 2017.

[6] M. Evers, L. Barnes, and M. Clark, "AMD next generation 'Zen 3' core," https://hc33.hotchips.org/program/, August 2021.

[7] L. Yang *et al.*, "Co-exploration of neural architectures and heterogeneous ASIC accelerator designs targeting multiple tasks," in *ACM/IEEE DAC*, 2020, pp. 1–6.

[8] Z. Shulin *et al.*, "3M-AI: A multi-task and multi-core virtualization framework for multi-FPGA AI systems in the cloud," in *ACM/SIGDA FPGA*, 2021, p. 228.

[9] Y. Xiang and H. Kim, "Pipelined data-parallel CPU/GPU scheduling for multi-DNN real-time inference," in *IEEE RTSS*, 2019, pp. 392–405.

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

[11] J. M. Joseph *et al.*, "Architecture, dataflow and physical design implications of 3D ICs for DNN accelerators," in *IEEE ISQED*, 2021, pp. 60–66.

[12] R. Mathur *et al.*, "Thermal-aware design space exploration of 3D systolic ML accelerators," *IEEE JXCDC*, vol. 7, no. 1, pp. 70–78, 2021.

[13] A. Samajdar *et al.*, "A systematic methodology for characterizing scalability of DNN accelerators using scale-sim," in *IEEE ISPASS*, 2020, pp. 58–68.

[14] J. Zhang and D. Tao, "Empowering things with intelligence: a survey of the progress, challenges, and opportunities in artificial intelligence of things," *IEEE IoTJ*, vol. 8, no. 10, pp. 7789–7817, 2020.

[15] H. P. Enterprise, "Cacti 7.0," 2019.

[16] Y.-H. Gong *et al.*, "Quantifying the impact of monolithic 3D (M3D) integration on L1 caches," *IEEE TETC*, vol. 9, no. 2, pp. 854–865, 2019.

[17] P. Ehrett *et al.*, "Analysis of microbump overheads for 2.5 d disintegrated design," *UMich. Ann Arbor Tech. Rep. CSE-TR-002-17*, 2017.

[18] S. K. Samal *et al.*, "Monolithic 3D IC vs. TSV-based 3D IC in 14nm FinFET technology," in *IEEE S3S*, 2016, pp. 1–2.

[19] K. Skadron *et al.*, "Temperature-aware microarchitecture," in *ACM ISCA*, 2003, pp. 2–13.

[20] A. Narayan *et al.*, "Prowaves: Proactive runtime wavelength selection for energy-efficient photonic nocs," *IEEE TCAD*, vol. 40, no. 10, pp. 2156–2169, 2020.

[21] S. D'souza and R. Rajkumar, "Cycletandem: Energy-saving scheduling for real-time systems with hardware accelerators," in *IEEE RTSS*, 2018, pp. 94–106.

[22] D. He *et al.*, "Network support for AR/VR and immersive video application: A survey," in *ICETE*. SciTePress, 01 2018, pp. 359–369.

[23] A. Heinig, R. Fischbach, and M. Dittrich, "Thermal analysis and optimization of 2.5D and 3D integrated systems with wide I/O memory," in *IEEE ITherm*, 2014, pp. 86–91.

[24] N. P. Jouppi *et al.*, "In-datacenter performance analysis of a tensor processing unit," in *ACM ISCA*, 2017, pp. 1–12.

[25] S. Cho *et al.*, "Mcdram v2: In-DRAM systolic array accelerator to address the large model problem in DNNs on the edge," *IEEE Access*, vol. 8, pp. 135 223–135 243, 2020.

[26] W. Liao, L. He, and K. M. Lepak, "Temperature and supply voltage aware performance and power modeling at microarchitecture level," *IEEE TCAD*, vol. 24, no. 7, pp. 1042–1053, 2005.