On Thermal Decoupling in Chiplet Systems

George Karfakis  
UCLA  
Los Angeles, CA, USA  
georgekarfakis@ucla.edu

Myriam Bouzidi  
Georgia Institute of Technology  
Atlanta, GA, USA  
mbouzidi3@gatech.edu

Yunhyeok Im  
Georgia Institute of Technology  
Atlanta, GA, USA  
yunhyeok.im@gatech.edu

Alexander Graening  
UCLA  
Los Angeles, CA, USA  
agraening@g.ucla.edu

Suresh K. Sitaraman  
Georgia Institute of Technology  
Atlanta, GA, USA  
suresh.sitaraman@me.gatech.edu

Puneet Gupta  
UCLA  
Los Angeles, CA, USA  
puneetg@ucla.edu

ABSTRACT
This paper investigates the thermal performance of tightly integrated heterogeneous chiplets in 2.5D systems, such as modern enterprise GPUs. In many such systems, thermally sensitive chiplets are thermally coupled to high power chiplets. For example, a GPU chiplet can heat up neighboring HBM chiplets resulting in performance degradation (effective bandwidth decreases by up to 27% for HBM3) due to increased refresh rates. This paper explores a variety of approaches to thermally decouple and isolate chiplets with a GPU+HBM case study. Our thermal simulations of a water-cooled 2.5D integrated GPU system indicate that the majority of thermal coupling occurs through the very low thermal resistance heat spreader, which renders conventional isolation approaches such as thermally-aware floorplanning not very effective. Furthermore, the effect of different interposer materials on thermal isolation is also comparatively limited. Instead, the most effective isolation approach is to introduce isolation materials within the heat spreader to effectively “split” the cooling between chiplets. We study the thermal trade-offs of such a split heat sink approach. To quantify the performance impacts of such a split, we develop a closed-loop workflow that incorporates thermal results when determining the runtime of deep learning workloads on GPU+HBM integrated systems. Our results demonstrate performance gains of up to 37% for some memory-bound workloads while utilizing the aforementioned ‘split’ heatspreader water-cooled design compared to baseline approaches.

1 INTRODUCTION
2.5D integration, where multiple silicon dies are integrated on a single interposer, has become a key technology to enable the latest high-performance systems-on-chip (SoC) and graphics processing units (GPUs) [21, 29]. This integration approach allows heterogeneous dies, such as logic, memory, and sensors, to be tightly integrated in a single package, providing benefits in terms of performance, power efficiency, and form factor [13].

However, a key challenge with 2.5D integration is managing the thermal dissipation of the multiple high-power chiplets within the package. Chiplet-based designs are especially sensitive to thermal issues because of the close proximity of heterogeneous dies and the shared heatspreader. Specifically, the simultaneous integration of high-power components such as compute dies along with temperature-sensitive devices such as High Bandwidth Memory (HBM) stacks, photonic transceivers, image sensors, and RF devices can lead to unacceptable degradation in performance and reliability. For example, HBM chiplets must aggressively increase their refresh frequency as the temperature increases, up to 4x the nominal refresh frequency [16], which, as shown later, has a significant performance impact.

Previous work has addressed the thermal challenges of 2.5D/3D integration primarily through thermal-driven floorplanning techniques [12, 5, 26, 3, 8, 20, 19, 6, 2]. These techniques show promise in reducing peak system temperatures; however, as this paper shows, they often necessitate larger interposer or package sizes, which can be impractical for large designs.

Various other thermal isolation strategies have also been explored. For example, thermal isolation air gaps within 3D stacks have been studied [37, 36], but this method is specific to 3D stacked chiplets and lacks generalizability to 2.5D systems.

[9] also contributes to this area by exploring differential cooling and varying interposer materials through specialized cooling manifolds in 2.5D systems. Our work builds on this by suggesting that, while different interposer materials can provide some thermal isolation, their effectiveness is limited in densely packed GPU+HBM configurations.

Additionally, [18] addresses thermal isolation using dielectric coolants within a microfluidic cooling framework, focusing on specific configurations and materials that may not be widely applicable.

Furthermore, the utilization of thermal isolation materials, such as aerogel, as a method of isolation has also been explored. [34] proposed a way to utilize aerogels to protect a heat-sensitive MEMS sensor from its processing electronics. In addition, [24] explored the use of aerogel as an isolation material within the interposer.

Although promising, this approach neglects to model the significant thermal coupling caused by the thick heat spreader, which, as we argue, is the primary medium for thermal coupling in 2.5D systems.

Lastly, [35] investigated thermal management using high conductivity materials to create new thermal paths between chiplets. Although this method is informative, our findings emphasize the importance of directly addressing thermal coupling pathways, particularly within the heat spreader, to fully mitigate thermal issues.

In this work, we compare and contrast these 2.5D thermal isolation approaches and evaluate their impact on system performance in the context of a GPU-HBM system. The key contributions of this work are as follows.

- With a carefully set thermal simulation environment, we show that the majority of inter-chiplet heat transfer occurs through the high thermal conductivity heat spreader.
rather than the interposer. As a result of this, we show that thermally-driven floorplanning approaches have limited benefits.

- We show that the emerging technique of inserting thermal isolators embedded within the heat spreader is far more effective in preventing heat transfer to temperature-sensitive chiplets.
- We develop a closed-loop tool flow that links thermal simulation [1] with performance simulation of GPU [23, 22] and HBM [25] and use it to demonstrate that thermal coupling in GPU-HBM interposer systems can cause up to 38% performance degradation for convolutional network workloads. This reduces to only 1%-3% with the insertion of thermal isolators.

The paper is organized as follows. Section 2 motivates the need for thermal isolation with an HBM example. Section 3 compares various thermal isolation approaches in 2.5D systems. Section 4 quantifies the performance impact of thermal coupling and isolation for a GPU-HBM system on convolutional and GEMM benchmarks. Finally, Section 5 concludes the paper.

2 THERMAL COUPLING PITFALLS: A COMPUTE EXAMPLE

Thermal coupling between chiplets is a concern for a variety of heterogeneously integrated systems. In this section, we consider the example of computing systems with High Bandwidth Memory (HBM) chiplets co-integrated with CPU/GPU. Many modern machine learning accelerators are organized as a high power compute chiplet placed next to several HBM chiplets on a silicon interposer. HBM is structured as stacks of Dynamic Random Access Memory (DRAM), which is a capacitative memory technology. The leakage of capacitors increases with increasing temperature, leading to a decrease in the retention time of memory cells when subjected to high temperatures [7]. Consequently, the frequency of row refresh operations at high temperatures must increase, as reported in the JEDEC specifications for both HBM2 and HBM3 [15],[16].

According to JEDEC, HBM refresh behavior consists of two manufacturer-dependent trip point temperatures, beyond which refresh rates must be doubled. Previous works have reported these trip temperatures lie between 75-95 degrees [14][33][32][31]. For the purposes of this analysis, it is assumed that the trip points are at 75 and 85 degrees C. This is shown in Table 1.

<table>
<thead>
<tr>
<th>Refresh Rate</th>
<th>Temperature Trip Point (°C)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 x tREFI</td>
<td>&lt; 75</td>
</tr>
<tr>
<td>0.5 x tREFI</td>
<td>75</td>
</tr>
<tr>
<td>0.25 x tREFI</td>
<td>85</td>
</tr>
</tbody>
</table>

We use DRAMSim3 [25] to quantify the performance impact of HBM throttling by varying the temperature (and the subsequent refresh rate) while executing maximum-intensity HBM workloads (corresponding to 100% memory utilization). Table 2 shows that the HBM bandwidth decreases by 27%, read latency increases by 70%, and refresh energy increases by 400%.

<table>
<thead>
<tr>
<th>tREFI Ratio</th>
<th>Refresh Energy</th>
<th>Read Latency</th>
<th>Bandwidth</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.00</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>0.50</td>
<td>2.000</td>
<td>1.238</td>
<td>0.912</td>
</tr>
<tr>
<td>0.25</td>
<td>4.004</td>
<td>1.714</td>
<td>0.732</td>
</tr>
</tbody>
</table>

This lower effective bandwidth is potentially very limiting, especially when executing memory-bound workloads. Section 4 later shows the thermal coupling-driven slowdown caused in the application kernels for a GPU-HBM system that can exceed 30%. In the next section, we discuss and compare several well-known and emerging thermal isolation approaches which may be used to thermally decouple the heat-sensitive HBM from high-power compute.

3 A COMPARATIVE ANALYSIS OF THERMAL DECOUPLING APPROACHES

We investigate the impact of various thermal management and isolation approaches, including:

- Thermal floorplanning: Optimizing chiplet placement to minimize thermal coupling [26, 5, 3].
- Naive thermal 'spreading' of chiplets: Attempt to reduce thermal effects by spreading chiplets apart without any sophisticated thermal floorplanning algorithm. This aims to represent a more simplified version of thermal floorplanning. It preserves the topological relationship between chiplets while potentially reducing thermal coupling.
- Interposer material: Using different interposer materials to minimize thermal coupling [17].
- Thermal isolators: An emerging approach in which thermal isolators are introduced in the heatspreader and possibly in-between chiplets to enhance thermal isolation. [24, 36, 37].

3.1 Methodology and thermal simulation setup

This section will provide a comprehensive overview of the thermal simulation setup for the test model, which is a GPU + HBM system inspired by a Nvidia V100 GPU. The size, number and power of these chiplets are documented in Table 3.

<table>
<thead>
<tr>
<th>Chipset</th>
<th>GPU</th>
<th>HBM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Width [mm]</td>
<td>32</td>
<td>7.56</td>
</tr>
<tr>
<td>Height [mm]</td>
<td>25.5</td>
<td>11.49</td>
</tr>
<tr>
<td>Power [W]</td>
<td>202</td>
<td>12</td>
</tr>
<tr>
<td>Number of Chiplets</td>
<td>1</td>
<td>4</td>
</tr>
</tbody>
</table>

Figure 1 shows the top-down configuration of the V100-Inspired model.
The compute chiplet is modeled as monolithic silicon, while the somewhat complex HBM chiplet model is shown in Figure 2 (for a 4-tall HBM stack example). The rest of the paper assumes an 8-tall DRAM stack with a 100μm buffer I/O die that consumes 2.4W, 8 DRAM dies each 50μm thick that consume 1.2W, 20μm thick bonding layer between the DRAM dies and a 120μm thick bonding layer to the interposer. This model aligns closely with the configurations explored in [35]. The total power consumption of the HBM is set according to 7pJ/bit, consistent with the values reported in [27].

The HBM and compute chiplets are placed on an interposer (Silicon unless otherwise mentioned), which is placed on a FR-4 package/PCB that has modest backside cooling of 10W/mK. The PCB, along with the heat spreader, measures 266mm x 112mm, following the standard form factor for GPU cards. On top of the HBM and GPU dies, there is a 0.1mm thick layer of thermal interface material (TIM) and the copper heat spreader. The TIM conductivity is modeled at 100W/mK.

The heat spreader is a PCB-sized block of 15mm thick copper with 9 cooling channels built into it. The channels are fed by a water pump, with a water speed of 0.008m/s and total pump flow rate of 0.0023kg/s across all 9 channels.

### 3.2 Evaluating the efficacy of chiplet spacing for thermal isolation

There are multiple heat transfer pathways within an interposer system, as illustrated in Figure 3. We argue that although thermal floor planning and spreading techniques are potentially effective in reducing a system’s maximum temperature - primarily due to the enlarged thermal interface material (TIM) and copper indentation connecting to the heat spreader, resulting in a larger package size - their utility in achieving thermal isolation among different components is limited. This limitation arises from the unified heat spreader that covers the 2.5D chiplets, facilitating efficient lateral heat transfer through its high thermal conductivity copper.

Heat transfer between the GPU and HBM occurs through three distinct pathways. The first pathway is conduction through the interposer and the PCB/package. Additionally, the gap between the HBM and GPU permits heat exchange through minimal natural convection. However, the primary pathway is mediated by the thermal interface material (TIM) and the copper heat spreader. Among these, the thermal conductance of the heat spreader is notably superior, as its primary function is to dissipate heat from the components. In scenarios where there is a substantial temperature gradient between the devices, this attribute may prove counterproductive, as the heat transfer via the heat spreader exhibits the highest conductance and consequently the lowest thermal resistance. As indicated in Figure 3, the cross-sectional thermal conductance, a product of thermal conductivity and the cross-sectional thickness of the material, shows that the pathway that uses the TIM and the heat spreader is significantly more important than other pathways, notably exceeding the thermal conductivity of the pathways that cross the gap between the HBM and GPU and go through the interposer/PCB.

Crucially, separating chiplets by distance does not substantially affect this main heat path.

### 3.2.1 Evaluating the efficacy of chiplet spacing for thermal isolation

There are multiple heat transfer pathways within an interposer system, as illustrated in Figure 3. We argue that although thermal floor planning and spreading techniques are potentially effective in reducing a system’s maximum temperature - primarily due to the enlarged thermal interface material (TIM) and copper indentation connecting to the heat spreader, resulting in a larger package size - their utility in achieving thermal isolation among different components is limited. This limitation arises from the unified heat spreader that covers the 2.5D chiplets, facilitating efficient lateral heat transfer through its high thermal conductivity copper.

Heat transfer between the GPU and HBM occurs through three distinct pathways. The first pathway is conduction through the interposer and the PCB/package. Additionally, the gap between the HBM and GPU permits heat exchange through minimal natural convection. However, the primary pathway is mediated by the thermal interface material (TIM) and the copper heat spreader. Among these, the thermal conductance of the heat spreader is notably superior, as its primary function is to dissipate heat from the components. In scenarios where there is a substantial temperature gradient between the devices, this attribute may prove counterproductive, as the heat transfer via the heat spreader exhibits the highest conductance and consequently the lowest thermal resistance. As indicated in Figure 3, the cross-sectional thermal conductance, a product of thermal conductivity and the cross-sectional thickness of the material, shows that the pathway that uses the TIM and the heat spreader is significantly more important than other pathways, notably exceeding the thermal conductivity of the pathways that cross the gap between the HBM and GPU and go through the interposer/PCB.

Crucially, separating chiplets by distance does not substantially affect this main heat path.

The simple V100-inspired GPU+HBM chiplet design does not permit "thermal floorplanning", as its interposer is too small and the chiplets too densely packed to permit any nontrivial arrangements of chiplets. As such, we have focused on simply spreading the HBM chiplets apart from the GPU. Note that we are modeling the increase in communication link power due to the distance caused by spreading, following the analysis in [30]. For this setup, this means that up to 4.99W of extra power is required for the 10mm spreading case. Figure 4 shows that while distancing the chiplets from the base case of 0.5mm to up to 4mm, the temperature difference is minimal and never exceeds 3 degrees C for the GPU die or the
HBM chiplets. Meaningful differences are only prevalent at 10mm of separation, which is highly impractical due to the large package and interposer size increase it necessitates. As shown in Figure 4, to physically fit the chiplets while spreading, the underlying interposer had to increase in size by 11%, 22% and 55% for 2mm, 4mm and 10mm separation, respectively. This indicates that spreading is a poor strategy in thermal isolation, which corresponds to our assumptions about the importance of the top lateral thermal pathway. Thermal isolation through separation alone would require unreasonably large distancing between chiplets, which would result in manufacturing/packaging issues and cost increases (due to larger interposer size), as well as increased communication link power and decreased performance (due to longer communication links in the interposer).

Additionally, our analysis suggests that the thermal conductivity of the thin Thermal Interface Material (TIM) pad has a minimal impact on the overall results, given that the primary thermal conduction is mediated by the thick copper heatspreader. As shown in Figure 5, which presents the ΔTemperature between GPU and HBM, the thermal conductivity of TIM has a small effect on the results. From 20W/mK, which represents a conventional TIM [4], to 300W/mK, the difference in ΔTemperature never exceeded 0.75 degrees. The rest of the work uses TIM with 100W/mK conductivity.

3.3 Impact of interposer material on thermal coupling

We posit that, because of the excellent lateral top heat conductor, the effect of backside thermal spreading through the interposer will be relatively minimal. As such, switching the interposer material from Silicon to different, less conductive materials should not have a large effect. Figure 6 shows that less conductive interposer materials have a minimal effect on GPU or HBM temperatures. Switching from a Si interposer (thermal conductivity of 148W/mK) to a Glass (1.5W/mK) or FR-4 (0.3W/mK) only decreases peak HBM temperature by up to only 2.5 degrees and has a small effect on GPU temperature. This makes it a poor candidate approach to increase thermal isolation.

3.4 Effective thermal decoupling via thermal isolator insertion

We study an emerging alternative solution based on our central assumption that the main pathway for heat between chiplets remains the heat spreader on top of the system. We propose inserting thermal isolation materials such as aerogels or even FR-4 within the heat spreader to essentially "split" the heat spreader. This approach is similar to [24], but instead of limiting the isolating material to the interposer, we propose placing isolators within the copper heat spreader. In simpler systems where inter-chiplet distance is larger
On Thermal Decoupling in Chiplet Systems

than 1mm, this may be accomplished by literally splitting the heat sink into multiple separate heat spreaders.

With the addition of thermal isolators, as seen in Figure 7, the main lateral heat pathway is almost completely severed. The only thermal coupling paths are through the thin interposer/package or between the chiplets. As such, thermal coupling should be significantly reduced.

For this case study, we assume a 0.5mm-thick thermal isolator with thermal conductivity of an aerogel, 0.024 W/mK. This mimics an aerogel-like material, such as the one proposed in [24, 10]. This thermal isolator is placed throughout the entire heat spreader, essentially thermally "splitting" the heat spreader into 3 parts, as shown in Figures 8 and 9.

![Figure 8: Isolator setup. Not to scale, only indicative of isolator placement.](image)

![Figure 9: Isolator setup, top-down. The isolators cut across the heat spreader.](image)

Figure 10 shows the results of these thermal isolators in the 250W V100 inspired GPU model. In addition to effectively improving isolation, the proposed isolators allow for better control of fluid flow and cooling according to the needs of each isolated section of the heat spreader. As Figure 10 shows, changing relative flow rates in the baseline, non-isolated model has practically no effect on temperatures. This is because the highly conducting heat spreader effectively amortizes cooling across all fluid channels, so different flows per channel cannot affect local parts of the heatspreader. However, when isolation is used, this avenue of thermal control is opened to us. We utilize this to ensure that the GPU remains relatively cool even under the highest isolation conditions, by allocating 30% more cooling flow to its cooling channels compared to the memory channels.

From a baseline temperature of 85.1 degrees, surpassing both HBM "trip points", incorporating a thin isolator within the heat spreader effectively reduces the HBM temperature to 74.39 degrees. Consequently, this modification increases the temperature differential ($\Delta$Temperature) between the GPU and the HBM in the system from 7.75 to 28.95 degrees. This significant alteration underscores the impact of thermal isolation strategies on managing the thermal dynamics within integrated circuit systems.

3.5 Thermal decoupling approaches in a complex heterogeneously integrated system

To ensure that our results are general, we conducted a similar battery of tests on a heterogeneously integrated complex (HI) system inspired by TAP2.5D [26], shown in Figure 12a. This system is much larger and allows for more sophisticated thermal isolation approaches.

![Figure 10: Baseline vs thermal isolator results](image)

Table 4: Complex HI System Specifications [26]

<table>
<thead>
<tr>
<th>Chiplet</th>
<th>CPU</th>
<th>GPU</th>
<th>HBM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Widths [mm]</td>
<td>12</td>
<td>18.2</td>
<td>7.75</td>
</tr>
<tr>
<td>Height [mm]</td>
<td>12</td>
<td>18.2</td>
<td>11.87</td>
</tr>
<tr>
<td>Power [W]</td>
<td>105</td>
<td>295</td>
<td>20</td>
</tr>
<tr>
<td>Number of Chiplets</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>

Table 4 shows the detailed setup of this new system. For this system, the overall cooling setup remains the same, but pump flow rates are increased to 0.5m/s or 0.072kg/s across all 9 channels to handle the increased total power.

In this work, in addition to the baseline model, we consider a "naively spread" model, in which the chiplets are spread out from each other as much as possible, shown in Figure 12e, and a "thermally floorplanned" model, in which the chiplets are spread around in a manner proposed by TAP2.5D [26] to minimize the overall peak temperature, shown in Figure 12f. Finally, we also
consider the thermal isolator case, as shown in Figure 12b. It is important to note that both the spread and thermally floorplanned models utilize an interposer area that is larger by 38.4% [26]. This is, for most designs, impractical. Furthermore, the complicated structure of this system does not permit us to scale the interposer to "split" the heat spreader and requires a simplified simulation model that, for example, ignores I/O power. Nevertheless, this is an interesting comparison case compared to the simple GPU setup we showcased above.

Complex HI model does not have a clean separation between memory water cooling channels and compute water cooling channels.

Figure 12b shows where these isolators have been placed. The exact placement of the isolators is an important hyperparameter, and optimizing it is beyond the scope of this paper. As such, this is a sample placement.

Figure 11 shows the results of these thermal isolators on the Complex HI setup. As shown in Figures 12 (b), (c), and (d), three thermal isolator setups are examined.

- In setup (b), thermal isolators are placed to shield the HBMs from compute die heating. In this setup, with minimal compute heating and minimal increase in peak temperature, the temperature of the HBMs once again drops below the first trip point, increasing performance. This indicates that even with no increase in interposer area, thermal isolators can shield the HBM chiplets from heating effects.
- In setup (d), thermal isolators are instead placed to reduce the peak temperature between the compute chiplets. In addition, the size of the interposer is increased by 18%, half of what thermal floorplanning and naive spreading require, and the chips are spread modestly. HBM and peak compute temperatures are both dramatically decreased, maximizing performance. This shows how the increased TIM area results in lower vertical thermal resistance, which, coupled with isolators, is a very effective thermal management approach.
- In setup (c), we also present a version of setup (d) without the increased interposer area. This results in a moderate decrease in both HBM and compute temperature. Due to the fact that the TIM and interposer are limited in size, isolators with no spreading are of limited effectiveness when seeking to minimize peak temperature in this system.

4 PERFORMANCE IMPACT OF THERMAL COUPLING

In Section 2, we demonstrate how thermal heating of HBM chiplets can cause them to exceed their thermal design trip points, which has severe effects on their performance. In Section 3, we have showcased how thermal isolators can be used to isolate and shield HBM and other heat-sensitive chiplets from power-hungry compute dies in a 2.5D integration environment, preventing this effect.

We shall now proceed to evaluate the potential impacts of better thermal decoupling during the execution of AI workloads, illustrating the efficacy of the proposed thermal isolator-in-heatspreader design in mitigating these impacts.

4.1 Methodology and setup

The chosen platform for this analysis is a GPU-HBM system on a silicon interposer, similar to the Nvidia Titan V GPU. Notably, the Nvidia Titan V employs the identical package architecture and maintains the same Thermal Design Power (TDP) as the GPU+HBM system highlighted in previous results. This allows us to apply the thermal data presented in Section 3 to our selected system.

In order to quantify the exact performance impact of this thermally-driven throttling of compute and memory on real workloads, we utilized a workflow that incorporates Icepak [1], DRAMSim3 [25],
On Thermal Decoupling in Chiplet Systems

Figure 12: Comparison of five different Complex HI model setups highlighting various configurations and their thermal management strategies.

Figure 13: Performance workflow

First, we utilize a thermal simulator, such as Ansys Icepak, to create a table of GPU power vs. temperature for a given GPU system and heatspreader configuration that we wish to simulate. Two example tables (baseline / isolators) for the Nvidia TitanV GPU are shown in Tables 5 and 6.

Table 5: Baseline Power-Temperature table

<table>
<thead>
<tr>
<th>Power [W]</th>
<th>GPU [°C]</th>
<th>HBM [°C]</th>
</tr>
</thead>
<tbody>
<tr>
<td>202</td>
<td>92.97</td>
<td>85.10</td>
</tr>
<tr>
<td>168</td>
<td>86.14</td>
<td>80.68</td>
</tr>
<tr>
<td>134</td>
<td>79.21</td>
<td>76.57</td>
</tr>
</tbody>
</table>

Table 6: Thermal Isolator Power-Temperature table

<table>
<thead>
<tr>
<th>Power [W]</th>
<th>GPU [°C]</th>
<th>HBM [°C]</th>
</tr>
</thead>
<tbody>
<tr>
<td>202</td>
<td>103.34</td>
<td>74.39</td>
</tr>
<tr>
<td>168</td>
<td>94.07</td>
<td>72.15</td>
</tr>
<tr>
<td>134</td>
<td>84.76</td>
<td>69.91</td>
</tr>
</tbody>
</table>

Leveraging the data from these example Tables, as well as Tables 1 and 2 as detailed in Section 2, we establish a correlation between power consumption and performance metrics. Specifically, the following observations are made, assuming a maximum GPU temperature of 95 degrees:

- In the baseline heatspreader configuration, irrespective of the GPU power levels examined, thermal throttling of the HBM is anticipated due to increased refresh rates. However, the GPU temperatures remain below the threshold that would trigger thermal throttling.
- With the thermal isolator heatspreader configuration, there is no thermal throttling of the HBM across the range of GPU power values assessed. However, when the GPU power exceeds 168W, thermal throttling of the HBM is expected due to elevated temperatures.

To further our analysis, we initiate simulations of a specified workload using Accelsim [23]. These simulations generate a heatmap of multiple scenarios at different throttle points for the GPU (due to the decrease in clock frequency mediated by overheating) and the HBM (due to increased refresh rates when trip-points are exceeded).

4.2 Results

We employed the workflow described above to benchmark Polybench-conv2d [11], a convolutional filter application benchmark, and deepbench-inference, a general purpose GEMM workload [28].
Consider the example of polybench-conv2d, with an input size of 2048x2048. According to Accelwatch, the execution of this benchmark results in an average combined GPU+HBM power consumption of 254W, placing the system squarely within the throttling regime. Note that 254W includes the HBM power (48W, based on our assumptions), so the GPU die power is 206W.

In the scenario without thermal isolation, the system experiences maximal HBM throttling due to elevated memory temperatures. Conversely, GPU throttling is unnecessary, as its temperatures remain within safe operating limits. This significant degradation in memory bandwidth and increased memory latency adversely affect overall system performance, resulting in 32% slowdown.

In the scenario with thermal isolation, HBM throttling is completely prevented by maintaining lower memory temperatures. However, this isolation leads to elevated temperatures in the GPU, necessitating a 10% reduction in GPU clock speed to prevent overheating. This adjustment reduces the GPU power to 168W. This results in a slowdown of just 3% for this application kernel. This is shown in Figure 14.

Figure 14: Performance heatmap for polybench-Conv2D 2048x2048. Execution time is normalized.

The same runtime analysis was conducted for varying sizes of the polybench-conv2d benchmark, ranging from 2048x2048 to 8192x8192 input images and deepbench-inference-half 1760 x 7133 x 1760 General Matrix Multiplication (GEMM) benchmark. Notably, the deepbench GEMM benchmark consumes only 184W. Thus, the introduction of thermal isolators effectively reduces HBM temperatures below the first trip point, without incurring any performance drawbacks due to GPU throttling. The performance improvement from the insertion of isolators ranges from 9% to 37% on a variety of benchmarks as shown in Figure 15.

Figure 15: Performance results. Conv2D refers to the polybench-conv2d benchmark. Deepbench GEMM is 1760 x 7133 x 1760, half precision inference.

We propose and analyze the insertion of thermal isolators within the heat spreader as a viable thermal decoupling approach. We show that the targeted insertion of thermal isolators within a complex HI system can reduce peak temperature by 3.5% without any change in interposer area or 22.5% with an 18% increase in interposer area, while thermal floorplanning reduces peak temperature by only 9.98% with 38% interposer area overhead. We also showed that thermal isolator insertion in a GPU+HBM system can improve the performance of memory-bound machine learning workload kernels by as much as 37%.

This work opens up new avenues for thermal management and thermal-system co-optimization for 2.5D HI systems. The following are some examples of future work directions.

- The placement of the isolator within the system. For this example design, we have manually placed the isolators. However, the placement of isolators in a complex system is a challenging optimization problem with differing power consumption and temperature sensitivity profiles of different chiplets. In some thermally challenging high power systems, isolator placement may need to be jointly optimized with thermal floorplanning/spreading even if it comes with an area overhead.
- We assume a thin aerogel isolator of 0.5mm. Choice of isolator material and thickness can change lateral thermal coupling, and hence should be co-optimized with system architecture and workload. For instance, our preliminary results indicate that slightly thicker FR-4 thermal isolators can give comparable results, with cost, mechanical, and manufacturability benefits.
- This work has focused on microchannel liquid cooling setups. Physical thermal isolation in other cooling approaches such as air cooling needs to be studied.

5 DISCUSSION AND CONCLUSIONS

Our results clearly demonstrate that thermally aware chiplet floorplanning is very limited in its effectiveness and requires large (~4mm) separation between hot chiplets (e.g., GPU) and temperature-sensitive chiplets (e.g., HBM). This makes thermally-aware floorplanning impractical in most cases due to a communication performance penalty and increased interposer/package area. We show that the majority of thermal coupling occurs within the high thermal conductivity thick heatspreader rather than within the very low thermal conductivity thin interposer. Because of this, changing interposer materials is also ineffective in reducing thermal coupling.

We propose and analyze the insertion of thermal isolators within the heat spreader as a viable thermal decoupling approach. We show that the targeted insertion of thermal isolators within a complex HI system can reduce peak temperature by 3.5% without any change in interposer area or 22.5% with an 18% increase in interposer area, while thermal floorplanning reduces peak temperature by only 9.98% with 38% interposer area overhead. We also showed that thermal isolator insertion in a GPU+HBM system can improve the performance of memory-bound machine learning workload kernels by as much as 37%.

This work opens up new avenues for thermal management and thermal-system co-optimization for 2.5D HI systems. The following are some examples of future work directions:

- The placement of the isolator within the system. For this example design, we have manually placed the isolators. However, the placement of isolators in a complex system is a challenging optimization problem with differing power consumption and temperature sensitivity profiles of different chiplets. In some thermally challenging high power systems, isolator placement may need to be jointly optimized with thermal floorplanning/spreading even if it comes with an area overhead.
- We assume a thin aerogel isolator of 0.5mm. Choice of isolator material and thickness can change lateral thermal coupling, and hence should be co-optimized with system architecture and workload. For instance, our preliminary results indicate that slightly thicker FR-4 thermal isolators can give comparable results, with cost, mechanical, and manufacturability benefits.
- This work has focused on microchannel liquid cooling setups. Physical thermal isolation in other cooling approaches such as air cooling needs to be studied.

This paper highlights the critical importance of thermal-system co-optimization in tightly integrated 2.5D heterogeneous chiplet systems, particularly those that combine high-power GPU chiplets with thermally sensitive HBM chiplets. We have compared thermal
decoupling approaches in 2.5D systems and shown that previously assumed to work well approaches such as thermally-aware floorplanning are not very effective. We further propose and demonstrate the effectiveness of insertion of explicit thermal isolators to “split” the heatspreaders.

ACKNOWLEDGMENTS

This work was supported in part by CHIMES, one of the seven centers in JUMP 2.0, a Semiconductor Research Corporation (SRC) program, sponsored by DARPA.

REFERENCES