NNShim: Thermal Hotspots Simulation on ML Accelerators

Xuesi Chen, Margret Riegert, Daniel Ernst, Dave Werner, and Mark Hempstead

Electrical and Computer Engineering Department, Tufts University

1 Introduction

On-chip thermal hotspots have become an increasing concern for Machine Learning (ML) accelerators. There exists a need for software that can accurately predict the appearance of thermal hotspots during the architecture design stage and help designers mitigate hotspots in future ML accelerators. Our project, called NNShim, builds upon existing tools such as SCALE-Sim [4] for simulating data movement, and HotGauge [2] to generate temperature, power maps, and hotspot measurements. Our simulation results demonstrate that decreasing clock frequency; increasing aspect ratios; and applying row and column skipping can be effective ways to mitigate hotspots. Engineers can successfully use this tool to explore designs they would not otherwise be able to.

2 Background and Motivation

Existing ML accelerators can produce a significant amount of heat in the server (Google TPUv3 has a TDP of 450 W per chip), or, in the case of embedded systems, have very low thermal targets that must be met. Accurate thermal simulations of accelerator designs typically require the chip to already be designed, at which point its too late to make major design changes.

Systolic arrays are an array of small Multiply-Accumulate (MAC) units that specialize in matrix multiplication. These often suffer from hotspots, where heat is concentrated in certain areas and not others, creating a sharp temperature gradient in that area. This can cause problems, including decreasing performance and reducing device longevity. To increase the reliability and throughput of ML accelerators, and in order to meet the growing demand of ML applications in all industries, computer architects need to better consider power and thermal considerations in following generations of ML accelerators.

3 NNShim

We present NNShim, a power and thermal simulation tool for ML accelerators, with a focus on systolic arrays. The ultimate goal of NNShim is to assist designers in developing hotspot mitigations. Several simplifications and assumptions are made with the current version of the NNShim:

- Only systolic arrays and buffers are simulated and modeled
- Power is modeled at the granularity of a MAC unit and is assumed to be uniform within the unit

NNShim takes in an accelerator hardware configuration and a neural network to be inferenced and produces a thermal and a power trace. This grants users the flexibility to choose a specific ML accelerator floorplan and power consumption of both per operation of a MAC unit and per KB of buffer storage. The default NNShim power consumption setting is based on the work of Mathur [3] and Gemmini [1]. The key parameters are listed in Table 1.

<table>
<thead>
<tr>
<th>MAC area</th>
<th>MAC energy per op</th>
<th>buffer area</th>
<th>buffer read nrg</th>
<th>buffer write nrg</th>
<th>dataallow</th>
</tr>
</thead>
<tbody>
<tr>
<td>40 x 40 um</td>
<td>0.3 pJ</td>
<td>32502 um2/32 KB</td>
<td>1.1 pJ</td>
<td>1.5 pJ</td>
<td>OS</td>
</tr>
</tbody>
</table>

Table 1: Default Area and Energy consumption for systolic array and buffer simulated

NNShim interfaces with SCALE-Sim and the inputs of HotGauge. We modified SCALE-Sim so that it outputs activity maps (indicating parts of the systolic array that are active) and the total buffer read and write during a set sampling period. Activity maps are then used in conjunction with floorplan and power consumption information to generate the thermal and power trace through HotGauge.

There are several mitigation techniques NNShim provides for designers to use. One common technique is slowing down the clock. Another technique is row column skipping, which deactivates every other row or column of the systolic array (as seen in Figure 1b where the grey squares represent the deactivated MAC units), decreasing the total number of operating MAC units to 25% when both are used. The row column skipping technique is based on the assumption that all PEs have the ability to pass through data. It cools down systolic arrays by spreading power consumption and decreasing local power density while keeping the overall energy consumption the same. The last mitigation technique implemented is the aspect ratio, which changes the shape of the systolic array by powers of two.
Figure 1: A high level illustration of how NNShim bridges SCALE-Sim and HotGauge; a demonstration of row column skipping technique

4 Results

We are able to generate a power and temperature trace for each neural network and accelerator configuration selected. Some of the sample results can be found [here](#).

We studied the effect of clock frequency, aspect ratio, and row column skipping on the power consumption and temperature of the chip. For both of these results with Googlenet, Figure 3, and DLRM, Figure 2, above 1 GHz clock frequency the temperature doesn’t change much, but the hotspot severity tends to get worse. As expected, the temperature and thermal severity increased as clock frequency increased. However, changes in clock frequency and aspect ratio appears trivial in cooling down systolic arrays. In contrast, we found row column skipping to be the most effective cooling method, bringing down the max temperature on systolic array by 34% on average, but it also reduces the performance to 25%. Interestingly, for Googlenet, shown in Figure 3, turning on row and column skipping tends to make the hotspots worse in terms of thermal severity, which could be a result of increased maximum localized temperature difference. On the other hand, Increasing aspect ratio in general tends to decrease temperature but could increase or decrease thermal severity.

Figure 2: DLRM results showing aspect ratio being an effective mitigation technique in bringing down thermal severity. Row column skipping on the other hand, brings down the maximum temperature on systolic arrays significantly

Figure 3: Googlenet results showing aspect ratio of 2 creates worse thermal severity. Row column skipping technique on the other hand consistently shows its effectiveness on lowering the maximum systolic array temperature
References


