NNShim: Hotspots Mitigation for ML Accelerators

Margret Riegert, Daniel Ernst, Xuesi Chen

Problem Description

Machine learning applications have increased greatly in the past decade, both in the server and on the edge. In recent years there has been a move away from using GPGPUs to perform neural net inferences in favor of accelerators designed specifically for these workloads.

These 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.

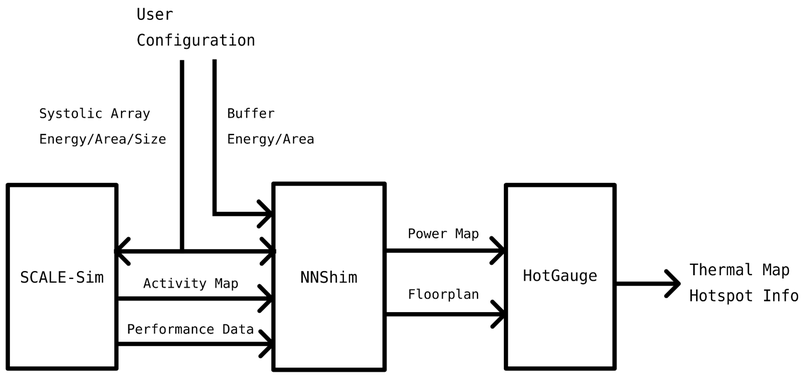
Accurate thermal simulations of accelerator designs typically require the chip to already be designed, at which point its too late to make major changes.

Solution / Implementation

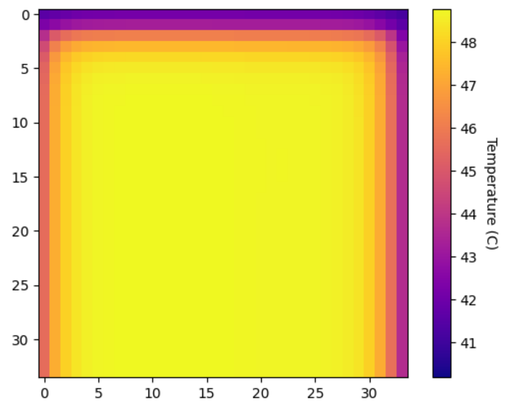
Investigations were done into various pre-existing systolic array simulators, and it was determined that SCALE-Sim was the best to work with for this project due to being written in Python and its permissive license.

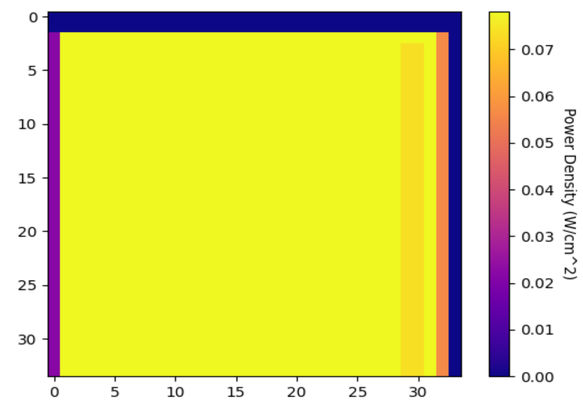
HotGauge is a thermal simulator that also has a Python interface, and was also chosen to work with this project due to it.

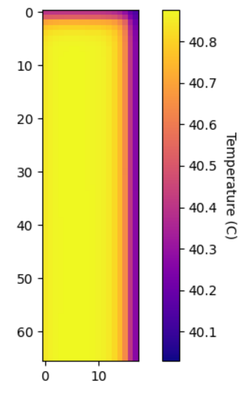
By combining these two with an intermediate program (NNShim), we are able to perform silicon-level thermal simulations using real ML workloads. A 3d array describing whether each part of the systolic array is active or not, called an activity map, was added as an export option to SCALE-Sim. This activity map is then used in conjunction with NNShim configuration information such as the size of the array, and the energy per operation to create a power map and floorplan. These are then handed off to HotGauge for thermal simulation.

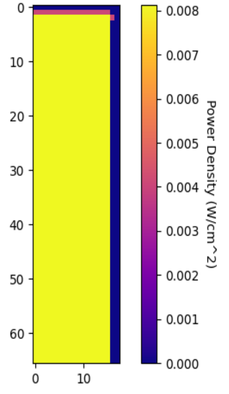


Clock frequency (Hz)	700 MHz - 1 GHz
PE energy per op	3 pJ
PE Size	40 um x 40 um
Buffer read/write energy per op	1.1 pJ / 1.5 pJ
SRAM buffer area	65004 um²
Systolic array size	32x32
Technology node	14nm









Results

We are able to generate a power and temperature trace for each neural networks and accelerator configurations selected. Above is two figures showing the temperature and power consumption of a 32X32 systolic array and its buffers on a certain timestamp.

We study the effect of clock frequency, aspect ratios and row column skipping on the power consumption and temperature of the chip. We find the row column skipping to be the most effective cooling method, but it also sacrifices the performance by half. The aspect ratio change the temperature on the chip only by a few degree Celsius.

Future Work

We are going to release the source code for NNShim publicly under an MIT license, as well as the modifications made to SCALE-Sim and HotGauge. Our hope is that NNShim is integrated into the HotGauge toolchain itself.

Future improvements to this could include modifying other simulators to work with NNShim instead of just SCALE-Sim, and optimizing SCALE-Sim for faster performance. As NNShim doesn't directly interact with SCALE-Sim, just through an activity map, any other simulator could be put in its place.