Study of Publicly Available TPU Architectures

Rathna Harsha

July 8, 2025

1 Wrapping Up the SoC

As we wind up the SoC, this last week we just take a look at how the neural network hardware is actually designed.

According to the problem statement, we had intermediate registers (or a common register bank) to save the activation values between each neuron and we had *one MAC block per unit*—which is very good in terms of computation speed but worse when we consider power consumption and area occupied by the hardware.

2 Few drawbacks of current model

Now let's leverage a few facts to our benefit:

1) We know that we aren't using up all the MAC blocks at a given time, so why can't we use a set of MAC blocks repeatedly over time?
Yes we can, but it would need planning—what MAC blocks to use for each time step.

2) We haven't optimized for matrix multiplications—yes, just like how we express them on paper, real–life hardware also has dedicated matrix multipliers to fast-track data inference and compute the results in one go. These are called *systolic array multipliers* (a whole new chapter to study deeper; heads-up, I am planning to do it for WIDS this winter)

3 Assignment

Now your work is to take a publicly available TPU hardware (Google's is available and you can find others as well) and make a **report on the architecture** of those hardware.

Your report should include, but is not limited to:

- How many parameters the architecture can handle;
- Power consumption and a comparison with a CPU;
- The number of operations per second it handles;
- Number of MAC blocks;
- What resources are available (memory hierarchy, interconnects, on-chip buffers, etc.);
- How the hardware accommodates training not just a fully connected (FC) neural network but also a CNN or RNN.
  (In the hardware you made it's limited to one FC neural network as it is under the hood of an FSM, so this can't be a unified hardware if your goal is to train other networks as well.)

4 Background on Systolic Arrays

A short overview of systolic array multipliers and why they are favored in modern tensor processing units (TPUs) can be included here. Feel free to expand.

5 Example architectures

Google TPU v1

Google TPU v2

Other Public TPU-like Accelerators