# Reading Reflection 4

Hao Deng

November 2023

## 1 Introduction

The reading materials cover the topic of the Roofline Model. The Roofline Model offers an approach to evaluate an application's performance on a machine by comparing it with the theoretical peak performance of the machine.

I've learned many basic ideas of this concept in AMATH 583 and I won't repeat them here. I want to mention something new.

There are several reasons that the performance is below the roofline.

1. Cache Bottlenecks
2. Fused Operations and Accelerators
3. Serialization
4. Lack of Parallelism
5. FPU Starvation

### 1.1 Cache bottlenecks

$$GFLOPS = min \begin{cases} \text{Peak GFLOP/s} \\ AI_{DRAM} * \text{DRAM GB/s} \\ AI_{L1} * \text{L1 GB/s} \\ AI_{L2} * \text{L2 GB/s} \end{cases}$$

This formula indicates that the performance is bound by the minimum performance among L1, L2, and DRAM.

### 1.2 Fused Operations and Accelerators

Deep learning applications are a mix Tensor, FP16, and FP32 instructions. Therefore, there is an ceiling on performance defined by the mix of instructions. It's average performance, and hence lower than the peak performance.

### 1.3 Serialization

Communication and computation are serialized, i.e. interleaving.

### 1.4 Lack of Parallelism

Perfect scaling is a vertical line in the roofline model.
Too many SMs(Streaming Multiprocessors) may cause cache capacity exhaustion, which leads to decrease in performance.

## 1.5 FPU Starvation

I didn't understand this idea at the very beginning. According to Shuai Lu, this indicates FPUs (Floating-point Unit) are not running at the peak performance because some of them are "starving", i.e. don't have data to process. This may occur when the applications are bounded by bandwidth.