week10实验记录

zxp

May 4, 2024

1 environment

cpu:Intel(R) Xeon(R) Gold 6330 CPU @ 2.00GHz (使用时申请了56个核心并且使用独占指令)

gpu:rtx3090(使用时申请了一块)

System:CentOS7

Compiler:9.5

2 code

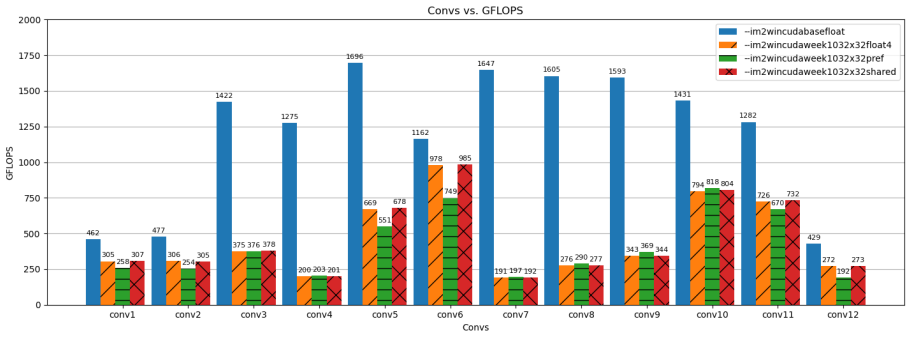

Figure 1: GPU

3 Experiment

这周还是在负优化。如果完全不考虑im2win张量中计算output行的时候可以复用的元素，使用共享内存的时候可以一次不全部搬运，而是将数据裁成一小块一小块放入共享内存中间，这周在完全未优化的卷积上面使用了这种方式，但是效果十分差。
然后试了其他优化，向量化加载使用float4，数据预取使用两块不同的共享内存一块读一块写。效果也很差。

3.1 Analysis

使用了共享内存比不使用还差应该是分块分的不好所有线程块里面的线程同一时刻都需要用同一个共享内存使得有点冲突，float4和预取效果差应该是分块分的不好，也没有填充没有考虑除4后剩余部分的处理。这样写只有负面优化，下周直接用卢帅师兄用的隐式矩阵乘法。

4 Experiment2

nsight使用nsys profile 运行程序后会给出一个报告，但这个报告的指标不包括核函数的具体信息，只有核函数运行时间和搬运的显存大小，还有个指标理论占用率（每个线程束同时工作的线程）只有66%（应该就是需要读同一个共享内存导致停顿）。核函数需要使用Nsight Compute。Nsight Compute 还没试。