CS133 Lab 3 Report

Yanzhe Liu 504153440

Parallelization Strategy:

In this laboratory, I first measured the time spent on each step in the convolutional neural network in the sequential version. I found that the convolution step takes most the total time (over 98%), so I decided to parallelize this step. I created a work group size of 256 to parallelize the convolution step. The outermost loop of the convolution step is distributed to each thread. Local buffers are used to store values of weight to improve cache performances. I applied vectorization at the end to gain some speedup.

The expected communication overhead is the overhead involved in moving data such as Cin, weight, C.

Evaluation:

Work-item Size	1	128	256
Performance (GFlop/s)	3.14	~60	~70

Table 1. Performance with different experimental parameters

Challenge:

The most challenge part of this project is understanding OpenCL memory structure. The naïve parallelization without any optimization has very poor performance.