|  |  |
| --- | --- |
| **Name:** | *Qian Liyang* |
| **NetID:** | *3190110719* |
| **Section:** | *ZJUI* |

**ECE 408/CS483 Milestone 2 Report**

|  |
| --- |
| 1. Show output of rai running Mini-DNN on the basic GPU convolution implementation for batch size of 1k images. This can either be a screen capture or a text copy of the running output. Please do not show the build output. (The running output should be everything including and after the line "*Loading fashion-mnist data...Done*"). |
|  |
| 1. For the basic GPU implementation, list Op Times, whole program execution time, and accuracy for batch size of 100, 1k, and 10k images. |
| |  |  |  |  |  | | --- | --- | --- | --- | --- | | Batch Size | Op Time 1 | Op Time 2 | Total Execution Time | Accuracy | | 100 | *0.182508 ms* | *0.6552 ms* | *0m1.143s* | *0.86* | | 1000 | *1.69811ms* | *6.47096ms* | *0m9.645s* | *0.886* | | 10000 | *16.7434 ms* | *64.4371 ms* | *1m37.553s* | *0.8714* | |
| 1. List all the kernels that collectively consumed more than 90% of the kernel time and what percentage of the kernel time each kernel did consume (start with the kernel that consumed the most time, then list the next kernel, until you reach 90% or more). |
|  |
| 1. List all the CUDA API calls that collectively consumed more than 90% of the API time and what percentage of the API time each call did consume (start with the API call that consumed the most time, then list the next call, until you reach 90% or more). |
| More than 90% is cudaMecpy and cudaMalloc |
| 1. Explain the difference between kernels and CUDA API calls. Please give an example in your explanation for both. 2. CUDA API calls for data transferring is larger than the kernel data transferring. For example, in above figures, the average for cudaMemcpy is 1.5e9 ns but for kernel, the total running time is just 4.1e8, which is only 27% of cudaMemcpy. That’s because the gpu have larger bandwidth than the main memory for host on most machines. 3. Cuda API is called by the CPU and kernerl is called by GPU. The CUDA API interface in nsys shows the functions running on cpu, which are related to CUDA library, for example, the cudaMemcpy . The CUDA Kernel interface shows the functions running on GPU, for example, the conv\_forward\_kernel |
| 1. Show a screenshot of the GPU SOL utilization |
| *Note: the screen shot is for 10,000 data size*  *Layer 1:*    *Layer 2:* |