|  |  |
| --- | --- |
| **Name:** | *Aditya Gupta* |
| **NetID:** | *adityag5* |
| **Section:** | *AL1* |

**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*"). |
| *Text  Description automatically generated* |
| 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 | *0.32 ms* | *1.197 ms* | *1.169 s* | *0.86* | | 1000 | *3.08 ms* | *11.985 ms* | *9.941 s* | *0.886* | | 10000 | *30.44 ms* | *120.418 ms* | *36.117 s* | *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). |
| *conv\_forward\_kernel – 100 %* |
| 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). |
| cudaMemcpy – 72.4 %  cudaMalloc – 20.2 % |
| 1. Explain the difference between kernels and CUDA API calls. Please give an example in your explanation for both. |
| *CUDA API calls are calls made by the written code into the CUDA driver or runtime libraries whereas the kernel is the function executed on the GPU N times in parallel by N different threads which may contains CUDA API calls.*  *In this case the conv\_forward\_kernel is the kernel and functions like cudaMalloc and cudaMemcpy are CUDA API calls.* |
| 1. Show a screenshot of the GPU SOL utilization |
| Timeline  Description automatically generated  *Timeline  Description automatically generated* |