HPCA Programming Assignment 2023-2024

Optimizing Performance of Dilated Convolution

Kirteyman Singh Rajput (SR: 21760)

November 26, 2023

[Part B] Implement and optimize DC in CUDA (GPU)

Optimizing dilated convolution for a GPU using CUDA

1. **Memory allocation on the GPU:** Allocation of meemory on the GPU for the input, kernel, and output matrices via *cudamalloc* function.
2. **Copy data to the GPU:** copy the input and kernel matrices from the CPU to the GPU using *cudaMemcpy* function.
3. **CUDA kernel:** Creating a CUDA kernel function that will be executed in parallel by multiple threads on the GPU.
4. **Invoke the CUDA kernel:** Specifying the number of threads and blocks and Launch the CUDA kernel ¡¡¡numBlocks, blockSize¿¿¿ syntax.
5. **Copy results back to the CPU:** Copy the output matrix from the GPU back to the CPU using *cudaMemcpy* after the computation is done on the GPU.

GPU Code with BlockDim=16 x 16			
	Input Matrix: 4096 Kernel Matrix: 11 Output Matrix: 4086	Input Matrix: 8192 Kernel Matrix: 11 Output Matrix: 8182	Input Matrix: 16384 Kernel Matrix: 11 Output Matrix: 16374
Execution time (ms)	242.569	676.956	2153.42
Reference execution time (ms)	16684.2	66920.4	270780
CPU cycles	4,53,48,42,218	16,84,59,36,291	67,19,46,34,431
Instructions	9,59,96,56,383	37,31,77,44,106	1,48,54,85,44,789
Instructions per cycle	2.12	2.22	2.21
L1 dcache loads	3,20,90,46,272	12,51,54,72,767	49,93,73,92,812
L1 dcache load misses	1,83,47,106	5,66,46,464	21,74,70,366
L1 dcache stores	1,33,49,07,465	5,18,82,55,198	20,70,15,75,990
Branch misses	93,19,009	3,12,98,281	12,97,11,897
Page faults	89,223	3,34,665	13,17,069
Context switches	43	353	315

GPU Code with BlockDim=32 x 32			
	Input Matrix: 4096 Kernel Matrix: 11 Output Matrix: 4086	Input Matrix: 8192 Kernel Matrix: 11 Output Matrix: 8182	Input Matrix: 16384 Kernel Matrix: 11 Output Matrix: 16374
Execution time (ms)	186.632	433.222	1382.08
Reference execution time (ms)	16684.2	66920.4	270780
CPU cycles	4,54,56,80,660	16,84,58,37,988	65,83,95,79,616
Instructions	9,63,22,50,743	37,25,48,73,087	1,47,93,60,52,396
Instructions per cycle	2.12	2.21	2.25
L1 dcache loads	3,20,81,95,905	12,53,36,10,483	49,75,52,15,125
L1 dcache load misses	1,82,74,660	5,66,33,957	20,59,00,506
L1 dcache stores	1,33,35,99,300	5,21,00,19,325	20,60,33,25,406
Branch misses	86,05,738	3,16,23,277	12,08,69,230
Page faults	89,226	3,34,666	13,17,068
Context switches	53	630	148

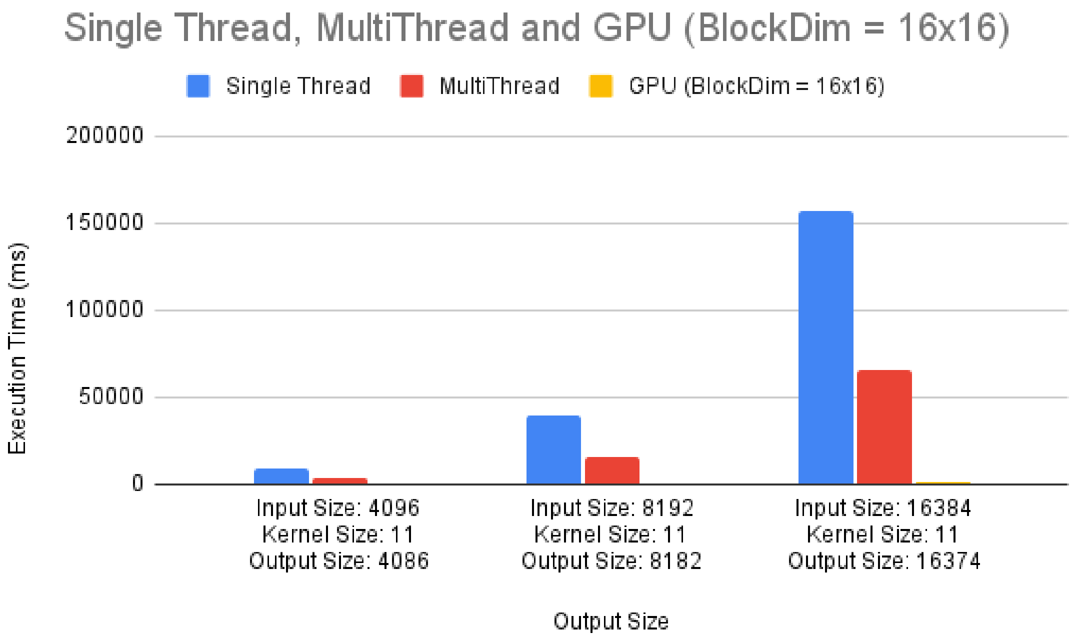

Figure 1: Single Thread vs Multithread Execution vs GPU Time(ms)

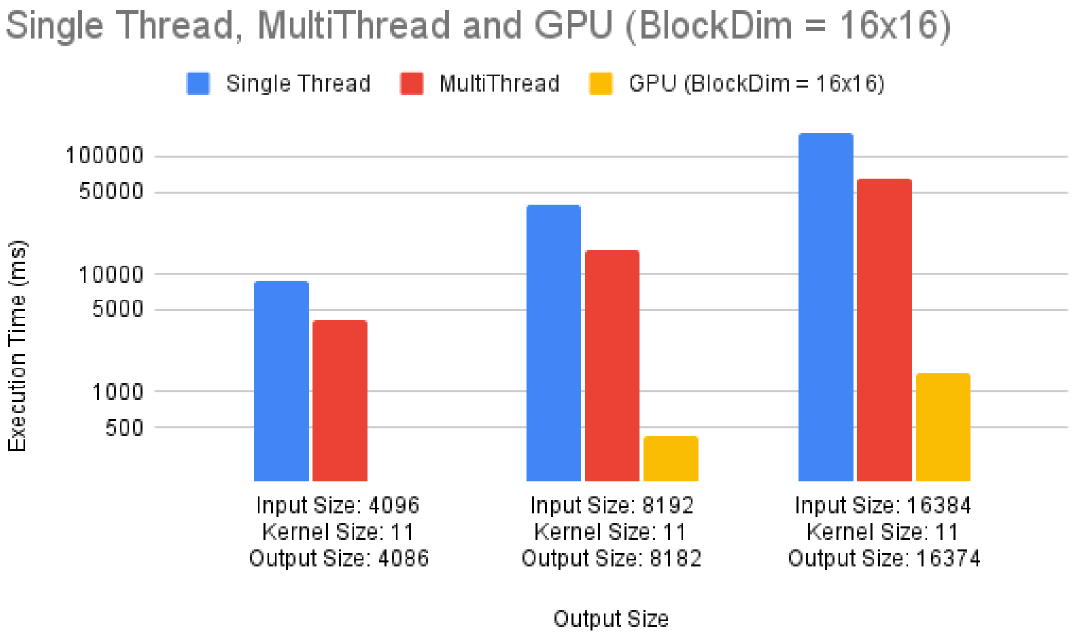

Figure 2: Logartithmic Graph: Single Thread vs Multithread Execution vs GPU Time(ms)