HPCA Programming Assignment 2023-2024

Optimizing Performance of Dilated Convolution

Kirteyman Singh Rajput (SR: 21760)

November 26, 2023

[Part A-I] Optimize single-threaded DC (CPU)

Dilated Convolution (DC):

Input:

A. Input Matrix of dimensions: Input_Row x Input _Column.

B. Kernel Matrix of dimensions: Kernel_Row x Kernel_Column.

Output:

An Output Matrix of dimensions:

(Input_Row - Kernel_Row + 1) x (Input_Column â Kernel_Column +1)

Unoptimized Code

Memory Access Optimization: Enhance memory access patterns to boost cache locality. Evaluate the reorganization of loops to enhance spatial locality, reducing the occurrence of cache misses.

Unoptimised Code			
	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) *(Reference)*	16684.2	66920.4	270780
CPU cycles	1,33,71,04,83,223	5,39,51,77,49,984	22,15,57,84,78,684
Instructions	2,63,79,54,50,938	10,57,67,54,90,975	42,35,42,37,09,979
Instructions per cycle	1.97	1.96	1.91
L1 dcache loads	1,18,42,63,78,965	4,74,86,59,74,734	19,01,83,62,03,888
L1 dcache load misses	1,47,19,73,030	3,55,47,66,163	13,99,81,42,817
L1 dcache stores	18,32,19,82,743	73,44,05,50,983	2,94,06,04,04,176
Branch misses	4,13,22,153	16,89,22,626	66,01,03,575
Page faults	81,745	3,27,182	13,09,583
Context switches	126	1,886	4,147

Optimized Code

Optimised Code			
	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)	8962.87	39140.2	156533
Reference execution time (ms)	16684.2	66920.4	270780
CPU cycles	1,03,61,32,78,806	4,23,60,47,48,688	17,56,42,31,57,924
Instructions	2,03,74,85,23,627	8,16,80,60,63,457	32,71,12,51,23,091
Instructions per cycle	1.97	1.93	1.86
L1 dcache loads	93,53,51,58,606	3,75,07,17,63,128	15,02,06,51,39,251
L1 dcache load misses	1,26,74,09,488	3,47,78,61,166	13,94,60,62,177
L1 dcache stores	16,87,60,83,129	67,66,88,62,566	2,70,96,46,69,157
Branch misses	4,12,11,544	16,67,62,954	66,06,81,727
Page faults	81,743	3,27,181	13,09,583
Context switches	308	385	1,678

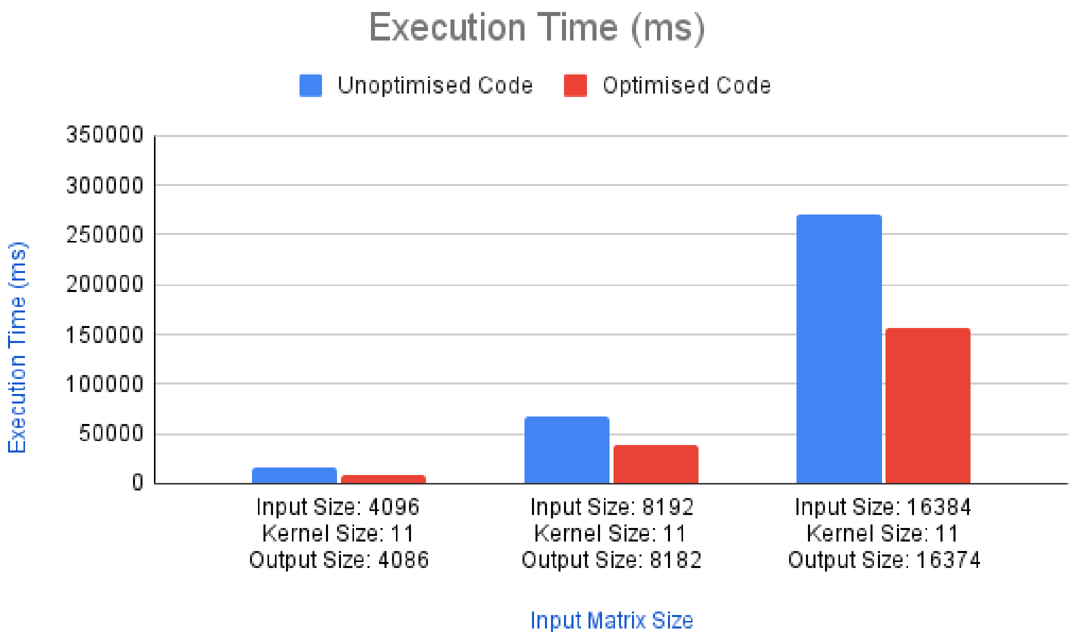

Figure 1: Single Thread Execution Time(ms): Optimised vs Unoptimised

[Part A-II] Implement and optimize multi-threaded DC (CPU)

MultiThreading:

In Multhread code, we distributed the equal number of rows of output matrix among the threads. Each thread will be responsible for a portion of the output rows. start_i determines the starting index of the rows for the thread i. and end_i determines the ending index (not inclusive) of the rows for the the thread i. start_i is means each thread starts at a different section of the output rows based on its ID and end_i ensures all rows are covered without overlap or missing rows.

MultiThread Code with Threads = 2			
	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)	9098.97	36161.5	154637
Reference execution time (ms)	16684.2	66920.4	270780
CPU cycles	1,42,05,20,38,591	5,69,02,46,67,009	23,47,79,47,13,677
Instructions	2,94,31,97,42,501	11,80,36,46,12,544	47,27,20,01,10,762
Instructions per cycle	2.07	2.07	2.01
L1 dcache loads	1,40,98,75,48,716	5,65,31,51,64,285	22,63,78,95,70,352
L1 dcache load misses	1,37,58,00,467	5,52,48,55,644	18,23,21,00,392
L1 dcache stores	18,30,10,40,257	73,37,20,99,656	2,93,79,53,36,716
Branch misses	4,17,88,189	19,50,43,059	76,82,01,769
Page faults	1,14,359	4,57,945	18,33,237 12,782
Context switches	609	4,561	12,782

MultiThread Code with Threads = 4			
	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)	4876.14	19779.9	77908.9
Reference execution time (ms)	16684.2	66920.4	270780
CPU cycles	1,45,29,58,26,128	5,88,47,07,71,869	23,50,36,12,09,804
Instructions	2,94,53,22,81,395	11,80,67,73,63,641	47,26,91,49,12,725
Instructions per cycle	2.03	2.01	2.01
L1 dcache loads	1,41,01,36,49,756	5,65,46,43,38,702	22,64,39,18,05,923
L1 dcache load misses	1,48,01,47,960	5,37,20,94,372	17,48,10,86,999
L1 dcache stores	18,30,40,03,457	73,39,16,34,809	2,93,96,21,61,976
Branch misses	5,06,58,991	20,08,33,505	78,15,62,375
Page faults	1,14,361	4,57,949	18,33,243
Context switches	446	1,559	5,635

MultiThread Code with Threads = 8			
	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)	4035.49	16143	65790.8
Reference execution time (ms)	16684.2	66920.4	270780
CPU cycles	1,93,95,87,16,550	7,76,75,37,35,601	31,65,63,70,56,559
Instructions	2,94,48,12,14,320	11,80,78,97,88,259	47,29,48,14,75,069
Instructions per cycle	1.52	1.52	1.49
L1 dcache loads	1,41,03,84,91,799	5,65,49,60,23,719	22,64,70,56,60,236
L1 dcache load misses	1,48,85,34,989	5,76,75,76,706	19,96,31,36,161
L1 dcache stores	18,32,54,07,664	73,46,21,01,566	2,94,25,88,84,179
Branch misses	4,81,01,495	18,79,78,108	79,91,73,558
Page faults	1,14,371	4,57,955	18,33,251
Context switches	2,098	7,522	34,684

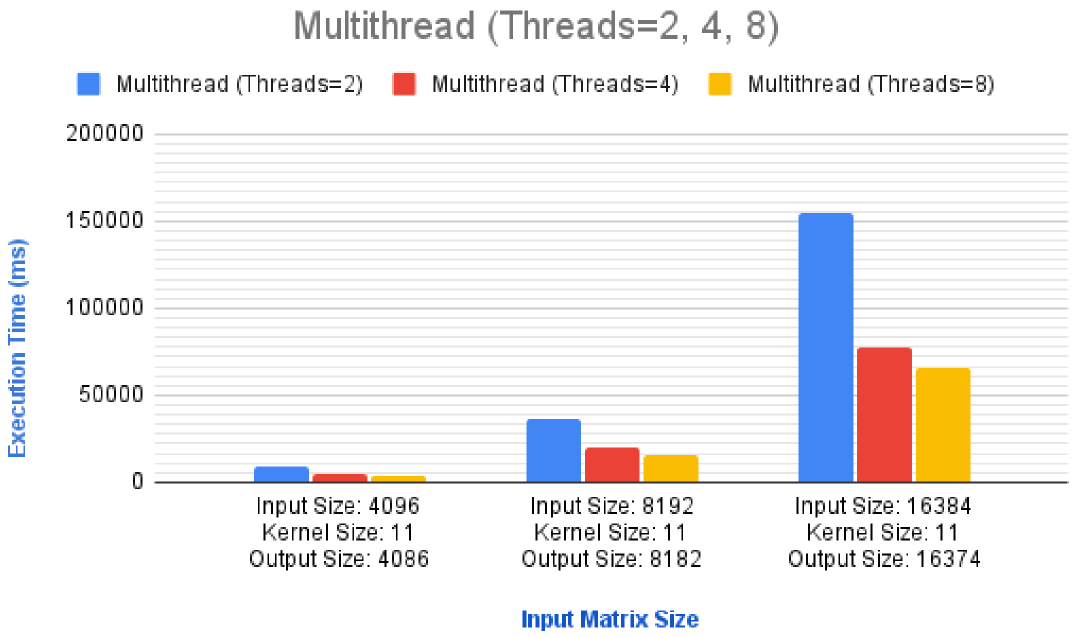

Figure 2: Multithread Thread Execution Time(ms)

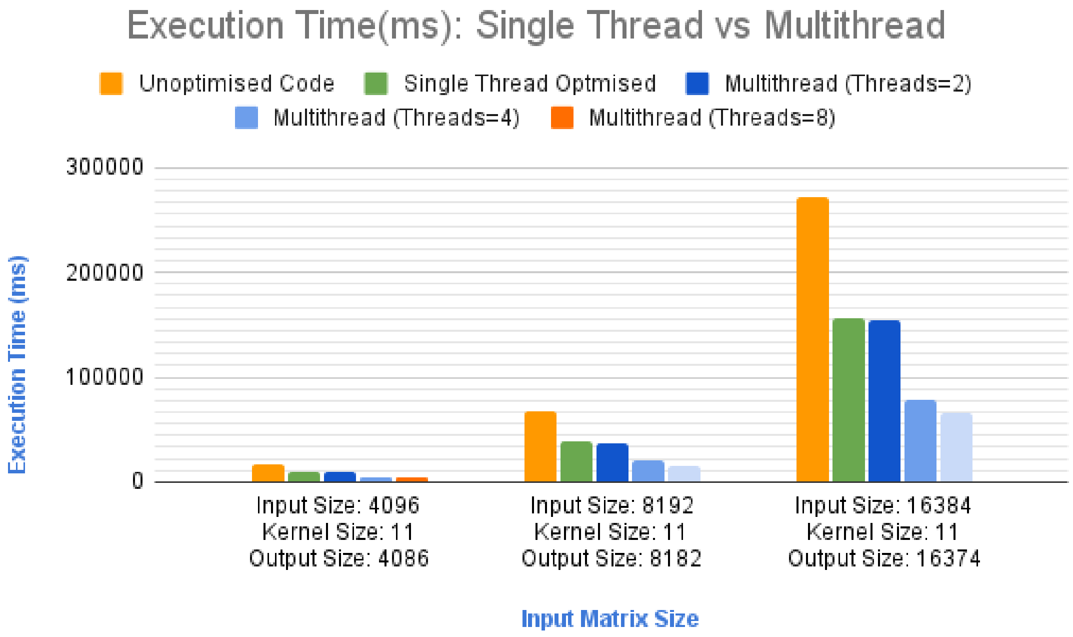

Figure 3: Single Thread vs Multithread Execution Time(ms)