Student Name: - ATHARVA DESHPANDE

Student Email: - deshpana@oregonstate.edu

Project No: - Project#6

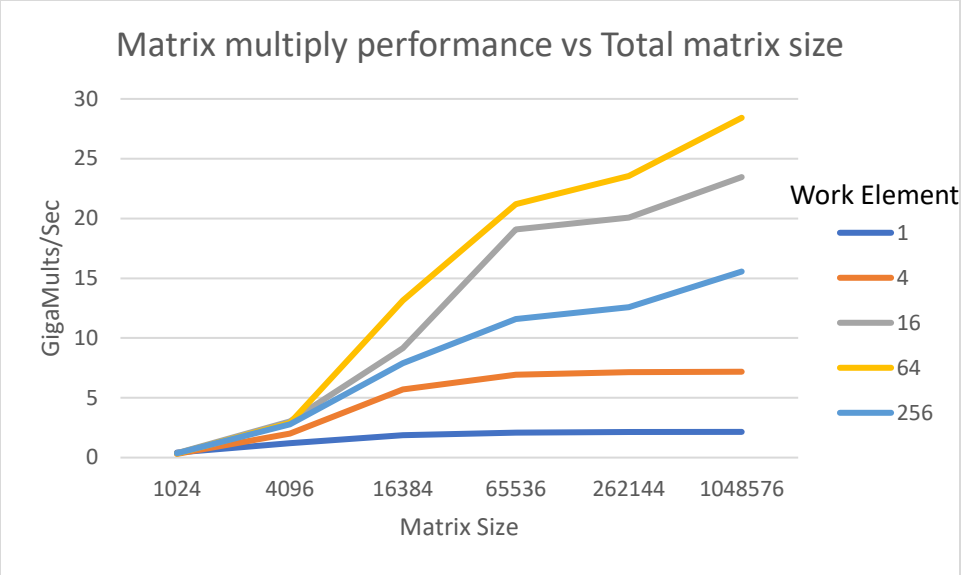
Project Name: - OpenCL Matrix Multiplication

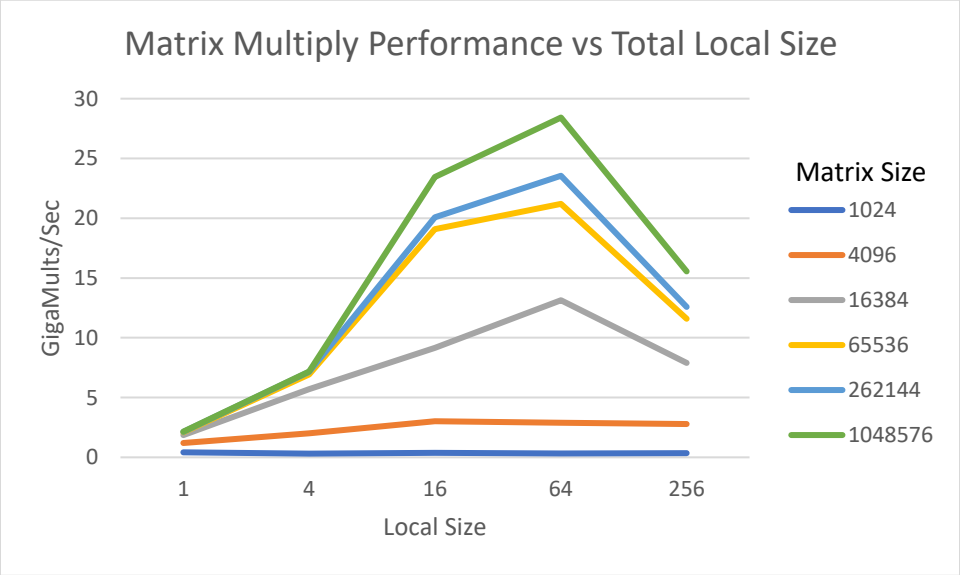
ATHARVA DESHPANDE CS575 deshpana@oregonstate.edu

1. What machine did you run this on?
   ➔ I ran this on the rabbit server which selected the vendor as NVIDIA for its GPU.

2. Show the table and graphs.
   ➔

Matrix Size	Work Element	GigaMults/Sec
1024	1	0.42
1024	4	0.31
1024	16	0.37
1024	64	0.33
1024	256	0.36
4096	1	1.2
4096	4	2
4096	16	3.02
4096	64	2.89
4096	256	2.78
16384	1	1.86
16384	4	5.71
16384	16	9.16
16384	64	13.14
16384	256	7.9
65536	1	2.09
65536	4	6.93
65536	16	19.08
65536	64	21.2
65536	256	11.58
262144	1	2.13
262144	4	7.15
262144	16	20.07
262144	64	23.55
262144	256	12.58
1048576	1	2.15
1048576	4	7.18
1048576	16	23.46
1048576	64	28.42
1048576	256	15.56





3. What patterns are you seeing in the performance curves? What difference does the size of the matrices make? What difference does the size of each workgroup make?

➔ In the first graph, we observe that increasing the matrix size leads to better performance. However, when it comes to the local size, performance stops improving after a certain point. Specifically, the highest performance is achieved with a matrix size of 10241024 and a local group size of 88, reaching 28.44 GigaMults/Sec.

In the second graph, we see that for matrix sizes of 3232 and 6464, performance remains the same regardless of the local size. Generally, increasing the matrix size improves performance,

except for the mentioned matrix sizes where performance stays consistent. The best performance is attained with a local work group size of 8*8, beyond which performance starts to decline, as depicted in the second graph.

4. Why do you think the patterns look this way?
   ➔ When working with larger matrices, the demands on processing power and memory bandwidth increase, which can create limitations and reduce performance. Additionally, factors like limited resources on the device or excessive memory usage by the workgroups can lead to contention and poorer performance. Another reason for using a local work size of around 8*8 is that it aligns well with the GPU's design and tends to be the most suitable choice.