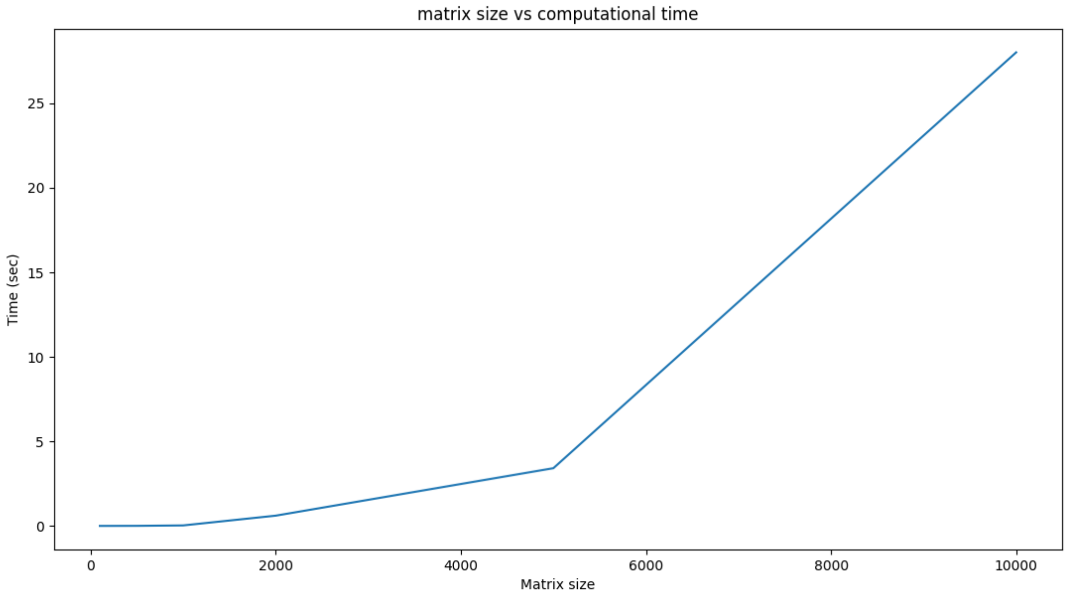
ME766-Assignment3 report

Dikshant 180040033

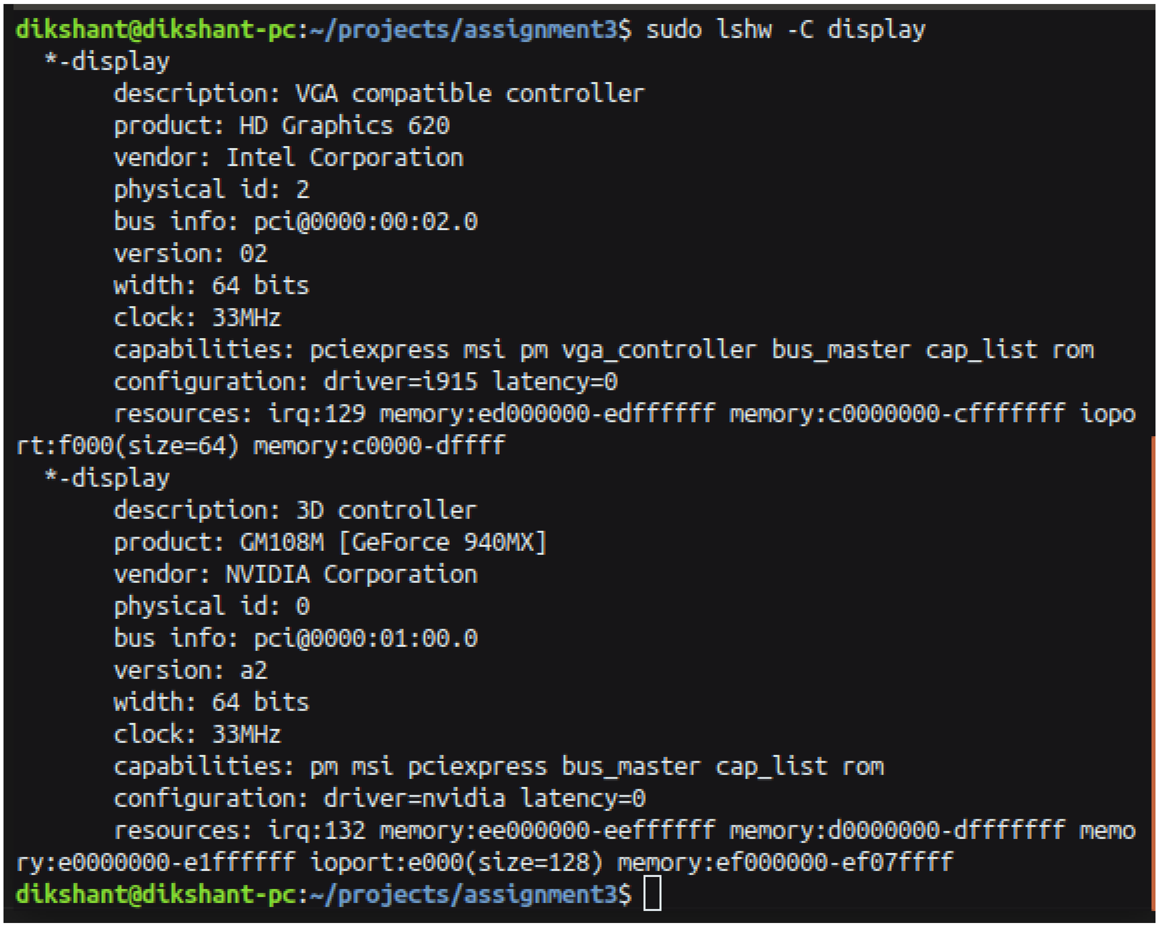
Q1: Create two matrices, A and B, each of size (N × N). Write a CUDA or OpenCL (choice is yours) for computing C = AB. Report the times taken for the codes. Vary the size of the problem from N = 100 . . . 10000.

Sol: I have done using CUDA. I have used tiled matrix multiplication for solving this task. It is an algorithm performed on GPUs due to the parallel nature of matrix multiplication. It is used to reduce global memory access by taking advantage of the shared memory on the GPU. Tiling can be seen to boost the execution efficiency of the kernel, basically to increase the "computation to memory ratio". One thread block computes one tile of the final matrix. One thread in the thread block adds one element of the tile. For example, let's take a 32 x 32 matrix consisting of four 16 x 16 tiles. Now, for its computation, 16x16 threads of each of these four-thread blocks can be created.

N	time1(ms)	time2(ms)	time3(ms)	Avg time(s)
100	0.364	0.357	0.360	.000360
500	4.892	4.909	4.855	.004885
1000	29.535	29.522	29.691	.029583
2000	202.015	202.681	202.612	.607308
5000	3416.389	3437.008	3391.049	3.414815
10000	28038.498	27588.292	28417.47	28.014753



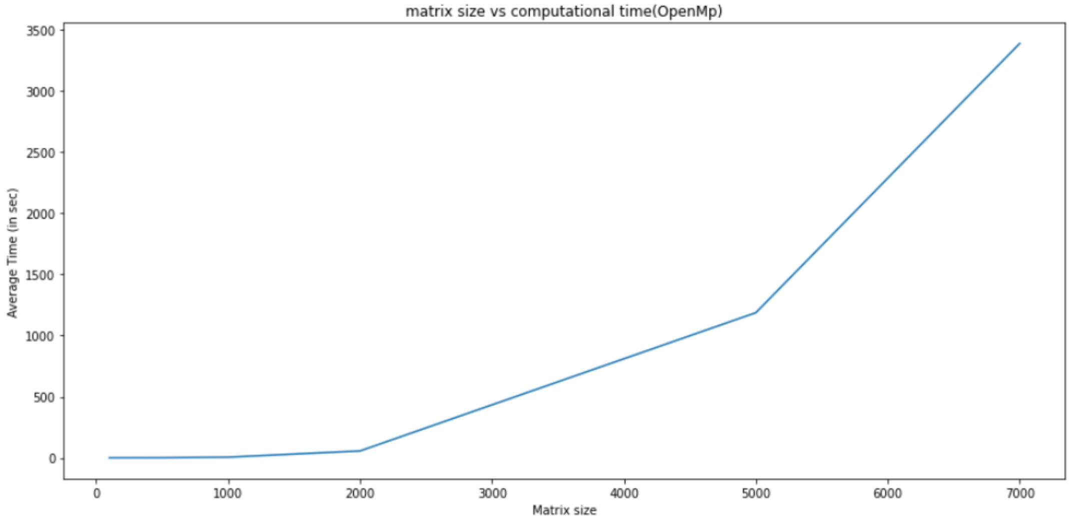
Specs of GPU: GeForce 940MX/PCIe/SSE2



```
dikshant@dikshant-pc:~/projects/assignment3$ sudo lshw -C display
  *-display
       description: VGA compatible controller
       product: HD Graphics 620
       vendor: Intel Corporation
       physical id: 2
       bus info: pci@0000:00:02.0
       version: 02
       width: 64 bits
       clock: 33MHz
       capabilities: pciexpress msi pm vga_controller bus_master cap_list rom
       configuration: driver=i915 latency=0
       resources: irq:129 memory:ed000000-edffffff memory:c0000000-cfffffff iopo
rt:f000(size=64) memory:c0000-dffff
  *-display
       description: 3D controller
       product: GM108M [GeForce 940MX]
       vendor: NVIDIA Corporation
       physical id: 0
       bus info: pci@0000:01:00.0
       version: a2
       width: 64 bits
       clock: 33MHz
       capabilities: pm msi pciexpress bus_master cap_list rom
       configuration: driver=nvidia latency=0
       resources: irq:132 memory:ee000000-eeffffff memory:d0000000-dfffffff memo
ry:e0000000-e1ffffff ioport:e000(size=128) memory:ef000000-ef07ffff
dikshant@dikshant-pc:~/projects/assignment3$ 
```

Q2: Choose A and B to be the same as HW 2.

Sol: I have used the same matrices, and as we can see, there's a significant difference between the computation time when we compute using CPU compared to GPU. A factor of around 125 reduced the timing for N=10000. (Attached graph is from the report of assignment 2)


matrix size vs computational time(OpenMp)
Average Time (in sec)
Matrix size