Assignment 4
CUDA C/C++ Acceleration

Arnav Goyal - 251244778

Graeme Watt - 251201502

April 8, 2024

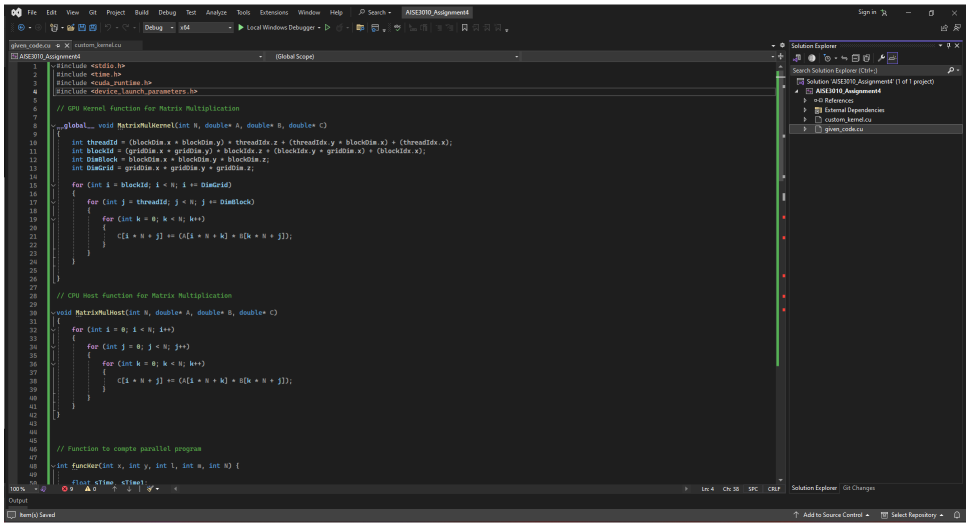
CUDA Environment

My CUDA environment was installed following the docs on NVIDIA's website. Here are my specs:

- Windows 11 Home 22H2
- GTX 1660-Ti 6 GB - Driver Version 552.12
- CUDA Compiler Driver 12.4
- Microsoft Visual Studio 2022 Community Edition (Version 17.9.5)

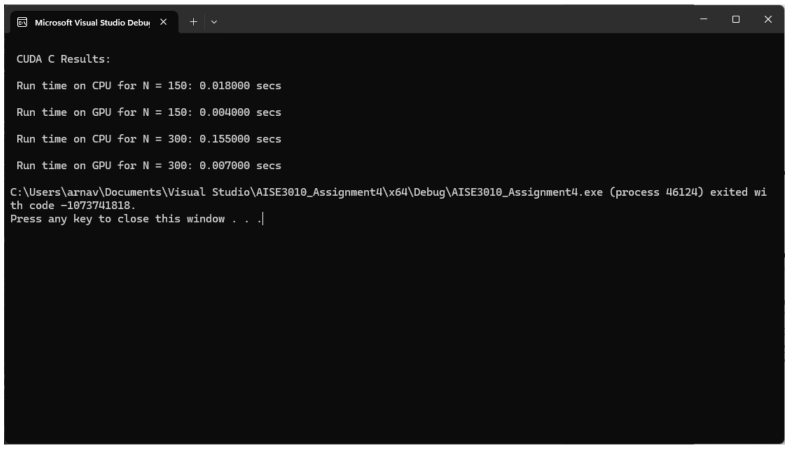
The Given Code

The `ipynb` file was uploaded onto colab, and the CUDA script was copied and pasted into my visual studio solution. It was built and ran without any modifications (other than including the neccessary `cuda_runtime.h` and `device_launch_parameters.h` header files), Im assuming that this is what *running smoothly means...* The given code produces the following console output.



```
#include <stdio.h>
#include <time.h>
#include <cuda_runtime.h>
#include <device_launch_parameters.h>

// GPU Kernel function for Matrix Multiplication

__global__ void MatrixMulKernel(int N, double* A, double* B, double* C)
{
    int threadId = (blockDim.x * blockDim.y) * threadIdx.z + (threadIdx.y * blockDim.x) + (threadIdx.x);
    int blockId = (gridDim.x * gridDim.y) * blockIdx.z + (blockIdx.y * gridDim.x) + (blockIdx.x);
    int DimBlock = blockDim.x * blockDim.y * blockDim.z;
    int DimGrid = gridDim.x * gridDim.y * gridDim.z;

    for (int i = blockId; i < N; i += DimGrid)
    {
        for (int j = threadId; j < N; j += DimBlock)
        {
            for (int k = 0; k < N; k++)
            {
                C[i * N + j] += (A[i * N + k] * B[k * N + j]);
            }
        }
    }

}

// CPU Host function for Matrix Multiplication

void MatrixMulHost(int N, double* A, double* B, double* C)
{
    for (int i = 0; i < N; i++)
    {
        for (int j = 0; j < N; j++)
        {
            for (int k = 0; k < N; k++)
            {
                C[i * N + j] += (A[i * N + k] * B[k * N + j]);
            }
        }
    }
}



// Function to compte parallel program

int funcKer(int x, int y, int l, int m, int N) {

    float sTime, sTime1;
```

Image 1: Given code in my CUDA environment



```
 CUDA C Results:

 Run time on CPU for N = 150: 0.018000 secs

 Run time on GPU for N = 150: 0.004000 secs

 Run time on CPU for N = 300: 0.155000 secs

 Run time on GPU for N = 300: 0.007000 secs

C:\Users\arnav\Documents\Visual Studio\AISE3010_Assignment4\x64\Debug\AISE3010_Assignment4.exe (process 46124) exited wi
th code -1073741818.
Press any key to close this window . . .
```

Image 2: Given code's output

The given code executes the kernel with the following parameters:

- a 3-dim grid size of 16, 16, 1
- a 3-dim thread block size of 16, 8, 1

The kernel uses these parameters to linearize the blocks and threads, to come up with a unique blockId and threadId for each parallel execution. The kernel then iterates over blocks, then iterates over threads) and performs the dot product of the row in A and the column in B and uses the result as an index of the matrix multiplication result C.

A big advantage of the code is that it is quite parallelized, and it basically calculates C through a whole 'rod' of the resultant 'cube' in each thread. This is better than performing a less parallelized matrix multiplication, for example writing a kernel that performs one matrix multiplication, and sequentially looping over it.

A big disadvantage of the code is that it is quite inefficient, it has time complexity of $O(n^3)$ (it has 3 nested for-loops), this makes it extremely inefficient and quite slow to larger and larger inputs. My solution is slightly better at $O(n^2)$

A Generalized Solution

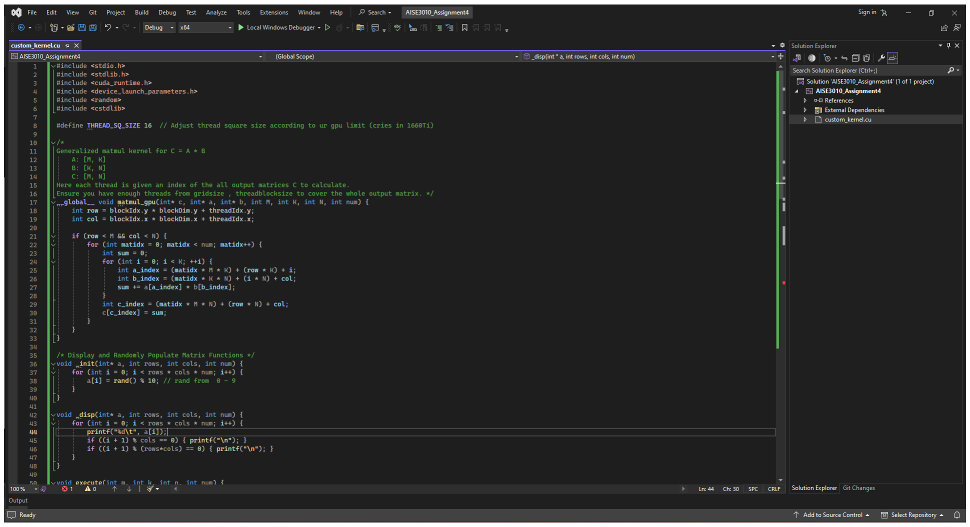
In order to generalize the code I started from scratch and created a new kernel, It works similarly to the given kernel but has some modifications, this kernel can be found in the `custom_kernel.cu` file. It works with any general matrix sizes for A and B that are compatible for the operation of multiplication. I tested its execution time for the following (given) scenarios:

i) At least 100 matrix multiplications of $A \times B$, for $A \in \mathbb{R}^{500\times500}$ and $B \in \mathbb{R}^{500\times400}$

ii) At least 5000 matrix multiplications of $A \times B$, for $A \in \mathbb{R}^{50\times20}$ and $B \in \mathbb{R}^{20\times50}$

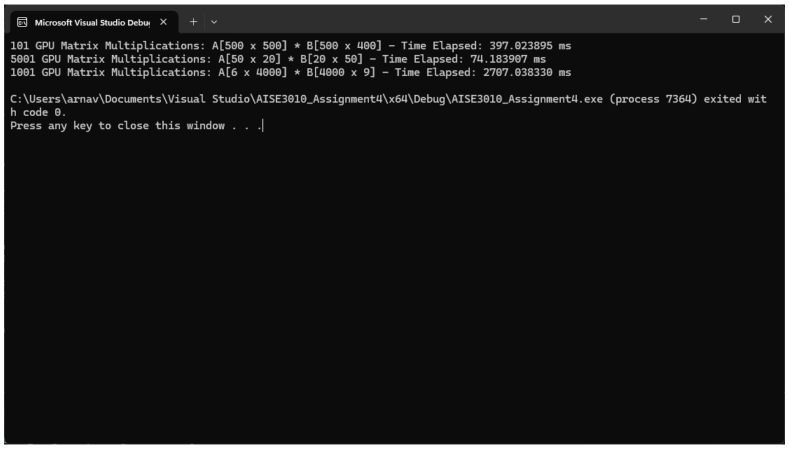
iii) At least 1000 matrix multiplications of $A \times B$, for $A \in \mathbb{R}^{6\times4000}$ and $B \in \mathbb{R}^{4000\times9}$

I also can examine the usage of my GPU with `Task Manager - Performance Monitor` and the `nvidia-smi -l 1` command. These two methods give the following results. Notice that the GPU utlization spikes to 100% from close to idle during build and output of my CUDA program.



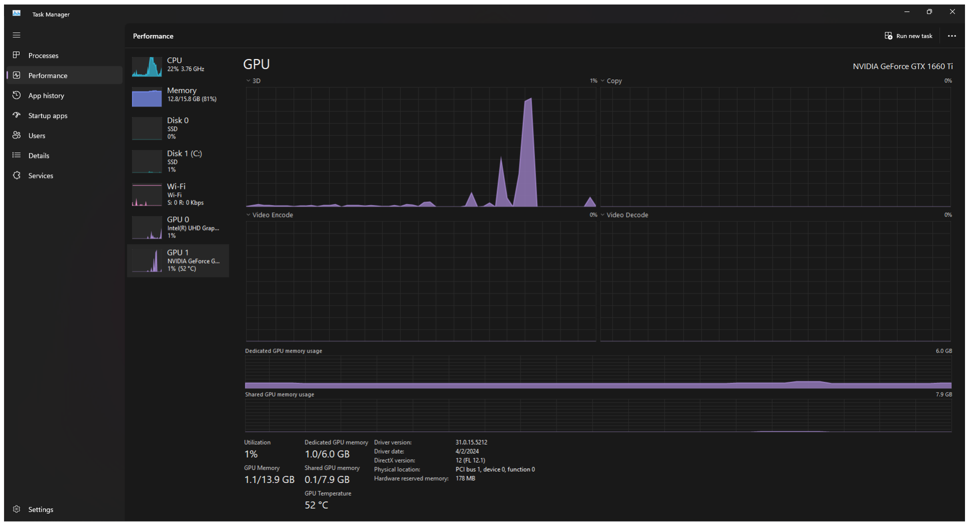
```
#include <stdio.h>
#include <stdlib.h>
#include <cuda_runtime.h>
#include <device_launch_parameters.h>
#include <random>
#include <cstdlib>

#define THREAD_SQ_SIZE 16  // Adjust thread square size according to ur gpu limit (cries in 1660Ti)

/*
Generalized matmul kernel for C = A * B
    A: [M, K]
    B: [K, N]
    C: [M, N]
Here each thread is given an index of the all output matrices C to calculate.
Ensure you have enough threads from gridsize , threadblocksize to cover the whole output matrix. */
__global__ void matmul_gpu(int* c, int* a, int* b, int M, int K, int N, int num) {
    int row = blockIdx.y * blockDim.y + threadIdx.y;
    int col = blockIdx.x * blockDim.x + threadIdx.x;

    if (row < M && col < N) {
        for (int matidx = 0; matidx < num; matidx++) {
            int sum = 0;
            for (int i = 0; i < K; ++i) {
                int a_index = (matidx * M * K) + (row * K) + i;
                int b_index = (matidx * K * N) + (i * N) + col;
                sum += a[a_index] * b[b_index];
            }
            int c_index = (matidx * M * N) + (row * N) + col;
            c[c_index] = sum;
        }
    }
}

/* Display and Randomly Populate Matrix Functions */
void _init(int* a, int rows, int cols, int num) {
    for (int i = 0; i < rows * cols * num; i++) {
        a[i] = rand() % 10; // rand from  0 - 9
    }
}

void _disp(int* a, int rows, int cols, int num) {
    for (int i = 0; i < rows * cols * num; i++) {
        printf("%d\t", a[i]);
        if ((i + 1) % cols == 0) { printf("\n"); }
        if ((i + 1) % (rows*cols) == 0) { printf("\n"); }
    }
}

void execute(int m, int k, int n, int num) {
```

Image 3: My CUDA program in the VS environment



```
101 GPU Matrix Multiplications: A[500 x 500] * B[500 x 400] - Time Elapsed: 397.023895 ms
5001 GPU Matrix Multiplications: A[50 x 20] * B[20 x 50] - Time Elapsed: 74.183907 ms
1001 GPU Matrix Multiplications: A[6 x 4000] * B[4000 x 9] - Time Elapsed: 2707.038330 ms

C:\Users\arnav\Documents\Visual Studio\AISE3010_Assignment4\x64\Debug\AISE3010_Assignment4.exe (process 7364) exited wit
h code 0.
Press any key to close this window . . .
```

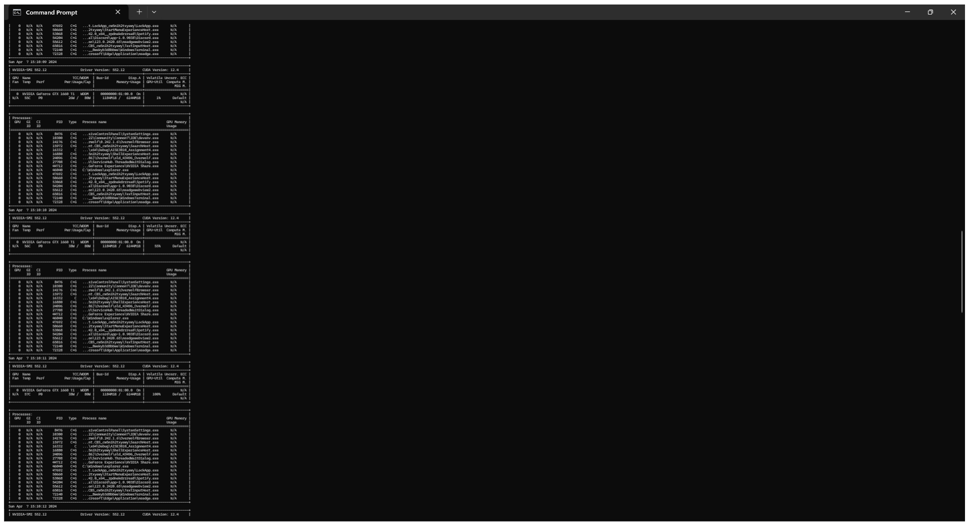
Image 4: output of my CUDA program on all three scenarios



Utilization
1%
GPU Memory
1.1/13.9 GB
Dedicated GPU memory
1.0/6.0 GB
Shared GPU memory
0.1/7.9 GB
GPU Temperature
52 °C

Driver version:	31.0.15.5212
Driver date:	4/2/2024
DirectX version:	12 (FL 12.1)
Physical location:	PCI bus 1, device 0, function 0
Hardware reserved memory:	178 MB

Image 5: GPU Utiliization spike in Task Manager's Performance Monitor



```
|    0   N/A  N/A     47692    C+G   ...t.LockApp_cw5n1h2txyewy\LockApp.exe      N/A      |
|    0   N/A  N/A     50660    C+G   ...2txyewy\StartMenuExperienceHost.exe      N/A      |
|    0   N/A  N/A     53068    C+G   ...42.0_x64__zpdnekdrzrea0\Spotify.exe      N/A      |
|    0   N/A  N/A     54204    C+G   ...al\Discord\app-1.0.9038\Discord.exe      N/A      |
|    0   N/A  N/A     55612    C+G   ...on\123.0.2420.65\msedgewebview2.exe      N/A      |
|    0   N/A  N/A     65016    C+G   ...CBS_cw5n1h2txyewy\TextInputHost.exe      N/A      |
|    0   N/A  N/A     72140    C+G   ...__8wekyb3d8bbwe\WindowsTerminal.exe      N/A      |
|    0   N/A  N/A     72328    C+G   ...crosoft\Edge\Application\msedge.exe      N/A      |
+---------------------------------------------------------------------------------------+
Sun Apr  7 15:10:09 2024
+---------------------------------------------------------------------------------------+
| NVIDIA-SMI 552.12                 Driver Version: 552.12         CUDA Version: 12.4     |
|-----------------------------------------+------------------------+----------------------+
| GPU  Name                     TCC/WDDM  | Bus-Id          Disp.A | Volatile Uncorr. ECC |
| Fan  Temp   Perf          Pwr:Usage/Cap |           Memory-Usage | GPU-Util  Compute M. |
|                                         |                        |               MIG M. |
|=========================================+========================+======================|
|   0  NVIDIA GeForce GTX 1660 Ti   WDDM  |   00000000:01:00.0  On |                  N/A |
| N/A   55C    P0             26W /   80W |    1184MiB /   6144MiB |      1%      Default |
|                                         |                        |                  N/A |
+-----------------------------------------+------------------------+----------------------+

+---------------------------------------------------------------------------------------+
| Processes:                                                                            |
|  GPU   GI   CI        PID   Type   Process name                            GPU Memory |
|        ID   ID                                                             Usage      |
|=======================================================================================|
|    0   N/A  N/A      8476    C+G   ...siveControlPanel\SystemSettings.exe      N/A      |
|    0   N/A  N/A     10300    C+G   ...22\Community\Common7\IDE\devenv.exe      N/A      |
|    0   N/A  N/A     14176    C+G   ...rwolf\0.242.1.6\OverwolfBrowser.exe      N/A      |
|    0   N/A  N/A     15972    C+G   ...nt.CBS_cw5n1h2txyewy\SearchHost.exe      N/A      |
|    0   N/A  N/A     16332      C   ...\x64\Debug\AISE3010_Assignment4.exe      N/A      |
|    0   N/A  N/A     16880    C+G   ...5n1h2txyewy\ShellExperienceHost.exe      N/A      |
|    0   N/A  N/A     24096    C+G   ...86)\Overwolf\old_43496_Overwolf.exe      N/A      |
|    0   N/A  N/A     27708    C+G   ...U\ServiceHub.ThreadedWaitDialog.exe      N/A      |
|    0   N/A  N/A     44712    C+G   ...GeForce Experience\NVIDIA Share.exe      N/A      |
|    0   N/A  N/A     46040    C+G   C:\Windows\explorer.exe                     N/A      |
|    0   N/A  N/A     47692    C+G   ...t.LockApp_cw5n1h2txyewy\LockApp.exe      N/A      |
|    0   N/A  N/A     50660    C+G   ...2txyewy\StartMenuExperienceHost.exe      N/A      |
|    0   N/A  N/A     53068    C+G   ...42.0_x64__zpdnekdrzrea0\Spotify.exe      N/A      |
|    0   N/A  N/A     54204    C+G   ...al\Discord\app-1.0.9038\Discord.exe      N/A      |
|    0   N/A  N/A     55612    C+G   ...on\123.0.2420.65\msedgewebview2.exe      N/A      |
|    0   N/A  N/A     65016    C+G   ...CBS_cw5n1h2txyewy\TextInputHost.exe      N/A      |
|    0   N/A  N/A     72140    C+G   ...__8wekyb3d8bbwe\WindowsTerminal.exe      N/A      |
|    0   N/A  N/A     72328    C+G   ...crosoft\Edge\Application\msedge.exe      N/A      |
+---------------------------------------------------------------------------------------+
Sun Apr  7 15:10:10 2024
+---------------------------------------------------------------------------------------+
| NVIDIA-SMI 552.12                 Driver Version: 552.12         CUDA Version: 12.4     |
|-----------------------------------------+------------------------+----------------------+
| GPU  Name                     TCC/WDDM  | Bus-Id          Disp.A | Volatile Uncorr. ECC |
| Fan  Temp   Perf          Pwr:Usage/Cap |           Memory-Usage | GPU-Util  Compute M. |
|                                         |                        |               MIG M. |
|=========================================+========================+======================|
|   0  NVIDIA GeForce GTX 1660 Ti   WDDM  |   00000000:01:00.0  On |                  N/A |
| N/A   56C    P0             38W /   80W |    1184MiB /   6144MiB |     55%      Default |
|                                         |                        |                  N/A |
+-----------------------------------------+------------------------+----------------------+

+---------------------------------------------------------------------------------------+
| Processes:                                                                            |
|  GPU   GI   CI        PID   Type   Process name                            GPU Memory |
|        ID   ID                                                             Usage      |
|=======================================================================================|
|    0   N/A  N/A      8476    C+G   ...siveControlPanel\SystemSettings.exe      N/A      |
|    0   N/A  N/A     10300    C+G   ...22\Community\Common7\IDE\devenv.exe      N/A      |
|    0   N/A  N/A     14176    C+G   ...rwolf\0.242.1.6\OverwolfBrowser.exe      N/A      |
|    0   N/A  N/A     15972    C+G   ...nt.CBS_cw5n1h2txyewy\SearchHost.exe      N/A      |
|    0   N/A  N/A     16332      C   ...\x64\Debug\AISE3010_Assignment4.exe      N/A      |
|    0   N/A  N/A     16880    C+G   ...5n1h2txyewy\ShellExperienceHost.exe      N/A      |
|    0   N/A  N/A     24096    C+G   ...86)\Overwolf\old_43496_Overwolf.exe      N/A      |
|    0   N/A  N/A     27708    C+G   ...U\ServiceHub.ThreadedWaitDialog.exe      N/A      |
|    0   N/A  N/A     44712    C+G   ...GeForce Experience\NVIDIA Share.exe      N/A      |
|    0   N/A  N/A     46040    C+G   C:\Windows\explorer.exe                     N/A      |
|    0   N/A  N/A     47692    C+G   ...t.LockApp_cw5n1h2txyewy\LockApp.exe      N/A      |
|    0   N/A  N/A     50660    C+G   ...2txyewy\StartMenuExperienceHost.exe      N/A      |
|    0   N/A  N/A     53068    C+G   ...42.0_x64__zpdnekdrzrea0\Spotify.exe      N/A      |
|    0   N/A  N/A     54204    C+G   ...al\Discord\app-1.0.9038\Discord.exe      N/A      |
|    0   N/A  N/A     55612    C+G   ...on\123.0.2420.65\msedgewebview2.exe      N/A      |
|    0   N/A  N/A     65016    C+G   ...CBS_cw5n1h2txyewy\TextInputHost.exe      N/A      |
|    0   N/A  N/A     72140    C+G   ...__8wekyb3d8bbwe\WindowsTerminal.exe      N/A      |
|    0   N/A  N/A     72328    C+G   ...crosoft\Edge\Application\msedge.exe      N/A      |
+---------------------------------------------------------------------------------------+
Sun Apr  7 15:10:11 2024
+---------------------------------------------------------------------------------------+
| NVIDIA-SMI 552.12                 Driver Version: 552.12         CUDA Version: 12.4     |
|-----------------------------------------+------------------------+----------------------+
| GPU  Name                     TCC/WDDM  | Bus-Id          Disp.A | Volatile Uncorr. ECC |
| Fan  Temp   Perf          Pwr:Usage/Cap |           Memory-Usage | GPU-Util  Compute M. |
|                                         |                        |               MIG M. |
|=========================================+========================+======================|
|   0  NVIDIA GeForce GTX 1660 Ti   WDDM  |   00000000:01:00.0  On |                  N/A |
| N/A   57C    P0             38W /   80W |    1184MiB /   6144MiB |    100%      Default |
|                                         |                        |                  N/A |
+-----------------------------------------+------------------------+----------------------+

+---------------------------------------------------------------------------------------+
| Processes:                                                                            |
|  GPU   GI   CI        PID   Type   Process name                            GPU Memory |
|        ID   ID                                                             Usage      |
|=======================================================================================|
|    0   N/A  N/A      8476    C+G   ...siveControlPanel\SystemSettings.exe      N/A      |
|    0   N/A  N/A     10300    C+G   ...22\Community\Common7\IDE\devenv.exe      N/A      |
|    0   N/A  N/A     14176    C+G   ...rwolf\0.242.1.6\OverwolfBrowser.exe      N/A      |
|    0   N/A  N/A     15972    C+G   ...nt.CBS_cw5n1h2txyewy\SearchHost.exe      N/A      |
|    0   N/A  N/A     16332      C   ...\x64\Debug\AISE3010_Assignment4.exe      N/A      |
|    0   N/A  N/A     16880    C+G   ...5n1h2txyewy\ShellExperienceHost.exe      N/A      |
|    0   N/A  N/A     24096    C+G   ...86)\Overwolf\old_43496_Overwolf.exe      N/A      |
|    0   N/A  N/A     27708    C+G   ...U\ServiceHub.ThreadedWaitDialog.exe      N/A      |
|    0   N/A  N/A     44712    C+G   ...GeForce Experience\NVIDIA Share.exe      N/A      |
|    0   N/A  N/A     46040    C+G   C:\Windows\explorer.exe                     N/A      |
|    0   N/A  N/A     47692    C+G   ...t.LockApp_cw5n1h2txyewy\LockApp.exe      N/A      |
|    0   N/A  N/A     50660    C+G   ...2txyewy\StartMenuExperienceHost.exe      N/A      |
|    0   N/A  N/A     53068    C+G   ...42.0_x64__zpdnekdrzrea0\Spotify.exe      N/A      |
|    0   N/A  N/A     54204    C+G   ...al\Discord\app-1.0.9038\Discord.exe      N/A      |
|    0   N/A  N/A     55612    C+G   ...on\123.0.2420.65\msedgewebview2.exe      N/A      |
|    0   N/A  N/A     65016    C+G   ...CBS_cw5n1h2txyewy\TextInputHost.exe      N/A      |
|    0   N/A  N/A     72140    C+G   ...__8wekyb3d8bbwe\WindowsTerminal.exe      N/A      |
|    0   N/A  N/A     72328    C+G   ...crosoft\Edge\Application\msedge.exe      N/A      |
+---------------------------------------------------------------------------------------+
Sun Apr  7 15:10:12 2024
+---------------------------------------------------------------------------------------+
| NVIDIA-SMI 552.12                 Driver Version: 552.12         CUDA Version: 12.4     |
```

Image 6: nvidia-smi command output looped every 1 second (you might need to zoom in)