

Programming Homework 6 - Report

Execution Times

	nStreams = 1	nStreams = 4	nStreams = 16
Approach 1	15.671264ms	12.742656ms	11.988800ms
Approach 2	15.721440ms	14.694400ms	14.693376ms

C[451][451]: 208,282,624

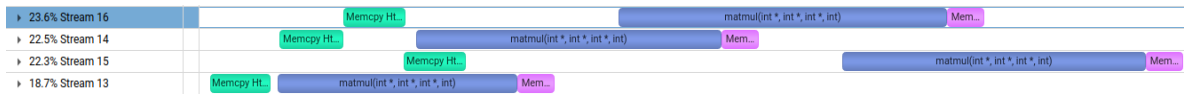
Platform:

GPU: RTX 3090

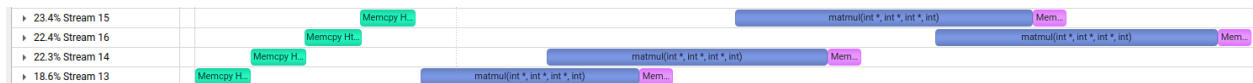
CUDA version: 12.6

Number of copy engines: 2

P1 Execution Timeline



P2 Execution Timeline



Observations

Execution times: As we can observe from the above execution timelines captured by nsight system, we observe that the kernels execute in parallel to the HtoD data transfer in approach 1. However, in approach 2, the kernels on each stream wait for the HtoD data transfer to complete across all streams before beginning execution, which causes longer execution times for approach 2.

I tried to find the reason for this behavior but could not find any information about the newer systems. [Nvidia Blog](#) quotes that the launch order is not important on newer systems however, we do not observe the same behavior.

Thus, I have put up a [post](#) on the NVIDIA forum to seek more information about the observed behaviour.

nStreams: Since in approach 1, kernel execution happens in parallel to the data transfer, we observe performance improvement as we increase n streams. However, increasing n to be too high will result in a performance drop due to the overhead of creating streams. Approach 2 also benefits a little from adding more streams as the kernel executions themselves are in parallel.

Output Screenshot

```
> ./p1  
N Streams: 4  
Time taken: 12.742656ms  
C[451][451]: 208282624
```

```
> ./p2  
N Streams: 4  
Time taken: 14.694400ms  
C[451][451]: 208282624
```