Assignment 5

2.d) What is the best performing value of block dim when n = 2_{15}?

Ans) block dim = 32 is the best performing value when n = 2^15.

2.e) Present the best runtime for n = 2_{15} from your HW04 matrix multiplication task (naive GPU implementation). Explain why the tiled approach performs better.

Ans) The best runtime for n =2^15 for HW4 is 787494 ms.

In the tiled based approach, the best runtime is 84851 ms.

The tiled approach is better because here the matrices data is being fetched from the shared memory which has significantly less latency compared to the data fetched from the unified memory which was the approach followed in HW4. Since, the data access is considerably faster in tiled approach due to shared memory hence the overall runtime of tiled approach is better.

2.f) Present the runtime for n = 2_{15} from HW02 (serial implementation mmul1) (or state that it goes beyond the Euler time limit). Explain why both GPU implementations perform better.

Ans) The runtime for n=2^15 for HW02 implementation goes beyond the Euler time limit. (Even a 1-hour limit was not enough)

The GPU implementations are faster because of the fine-grained parallelism allowed by the execution of multiple threads in the GPU architecture. Instead of the serialized execution of threads in the CPU implementation, the GPU implementations allows multiple threads to parallelly perform the execution operation for a large data set. Also, a higher bandwidth in the GPU architecture allows for fetching of large set of data from the unified memory and the execution is even faster with shared memory based tiled approach.