Assignment 4
GPU and CUDA

Arun Jayabalan (368048),
Chethan Shettar (368083),
Varun Gowtham (368053)

In this introductory lab on Graphical Processing Unit (GPU) and CUDA we are required to port a given single threaded regular C implementation able to generate a fractal, to GPU using CUDA kernel. The implementation concentrates on rendering the fractal using CUDA and write it into a PPM file.

For this assignment we have used GPGPU-Sim (www.gpgpu-sim.org) which provides simulation models of contemporary GPUs running on CUDA.

After simulating the CUDA kernel on the GPGPU-Sim, we obtain a log files containing details of execution and parameters.

We are required to simulate with different block sizes in 1-Dimensional and 2-Dimensional:
In 1-Dimensional blocks we simulate the following block sizes:
32, 256, 64, 128
In 2-Dimensional blocks we simulate the following block sizes:
8 x 8, 8 x 32, 32 x 8, 16 x 16

The following contains a brief report on the simulation results conducted on each blocks shown above:

2D Blocks

No.	BLOCK SIZE	Sim Cycle	Sim Instns	Instns/Cycle	Sim Time(sec)	Sim Rate(instn/sec)	Sim Rate(Cycles/Sec)
1	**4x4**	**316303**	**53053791**	**167.7**	**230**	**230668**	**1375**
2	**8x8**	**157386**	**53053791**	**337.09**	**190**	**279230**	**828**
3	**16x16**	**182592**	**53053791**	**290.5**	**199**	**266601**	**917**
4	**8x32**	**178548**	**53053791**	**297.14**	**219**	**242254**	**815**
5	**32x8**	**207069**	**53053791**	**256.11**	**217**	**244487**	**954**

No.	BLOCK SIZE	Stall	W0_Idle	W0_ScoreBoard	Total Waiting Time	SIMD Efficiency (%)
1	4x4	19509	110948	3216799	3347256	29.8786
2	8x8	309676	114935	579783	1004394	49.6835
3	16x16	547924	348270	222882	1119076	46.089
4	8x32	507084	336785	219699	1063568	49.6835
5	32x8	612396	383161	211395	1206952	41.0129

1D Blocks

No.	BLOCK SIZE	Sim Cycle	Sim Instns	Instns/Cycle	Sim Time(sec)	Sim Rate(instn/sec)	Sim Rate(Cycles/Sec)
1	1	2870097	52332895	18.2338	2909	17989	986
2	16	348066	52332895	150.3534	833	62824	417
3	32	230247	52332895	227.2902	354	147833	650
4	64	188672	52332895	277.3750	199	262979	948
5	128	190608	52332895	274.5577	187	279855	1019
6	256	201889	52332895	259.2162	192	272567	1051

No.	BLOCK SIZE	Stall	W0_Idle	W0_ScoreBoard	Total Waiting Time	SIMD Efficiency (%)
1	1	182271	105705	30011796	30299772	3.125
2	16	21815	130508	3516727	3669050	26.3683
3	32	13825	106580	2316973	2437378	40.7019
4	64	294676	117530	801070	1213276	40.7019
5	128	532502	342908	381010	1256420	40.7019
6	256	592590	493911	230867	1317368	40.7019

Observation:

- A Streaming Multiprocessor (SM) can contain 8 scalar cores
- Up to 8 Cores can run simultaneously
- When each core executes identical instruction set or sleeps
- Up to 32 threads may be scheduled at a time called a WARP.
- But max. 24 Warps may be active in a single SM.
- When the block size is more than 32, we observe the same SIMD efficiency, since all threads will be effectively utilized per Warp.
- Also when we have less number of Block size, we observe that the SIMD efficiency dips due to the fact that more threads will be in idle.
- For a block size of 1, the efficiency dips significantly. This can attributed to the fact that only one thread out of available 32 threads are utilized.
- In Case of 2-Dimensional blocks we observe that 8x8 and 8x32 has the highest efficiency of 49.6835%.
- In case of 1-Dimensional blocks we observe that starting from block size of 32, the efficiency remains same as 40.7019%.
- In case of 2-Dimensional the simulation we observe that the simulation cycles varies with different block sizes.
- In case of 1-Dimensional blocks we observe that the simulation cycles decreases as the Block size increases.