Ingara - Theory questions

1)
The host calculation took ~5.15 seconds to complete every time, while the CUDA one varied quite a bit, with the lowest I saw being 0.76 milliseconds, but with most being in the ballpark of 2-3 milliseconds. Copying the result from the GPU to the host typically took 3.5-5ms. So overall an improvement of three orders of magnitude. The speed-up is caused by the GPU's ability to perform the calculations for a significant number, if not all, of the pixels in parallel. In addition, the GPU has a fused multiply-add instruction which might have been used, further increasing performance.

2)
Using the cudaDeviceProp type I ascertained that I used a Tesla T4 with compute capability 7.5.

3)
SIMD stands for single instruction, multiple data, while SPMD stands for single program, multiple data. In SIMD the same operation is performed simultaneously on values from multiple registers, whereas in SPMD, several processors execute the same program simultaneously and typically operate on different data. According to NVIDIA, CUDA is SIMT, single instruction, multiple threads, which is a looser form of SIMD.