1. Timings for the host, naïve CUDA kernel, shared memory CUDA kernel and excessive memory copying case.

Program\ time per step (ms)	256	512	1024
the host	6.52012	6.675025	6.373175
naïve CUDA kernel	6.15415	6.245545	6.363265
shared memory CUDA kernel			
the excessive memory copying			

2. The GPU implementation is faster than the single threaded host code.

3. Haven’t done the shared memory kernel yet, but I guess it’s the excessive memcpy case.

4. Yes. Increasing the block dimension will increase the communication cost while multiple blocks access the same global memory. The latency from the additional blocks cannot be offset by the operation speed-up offered by the additional blocks.