** times listed are in milliseconds*

K	Threads	CPU Execution	GPU Execution	Host to Device	Device to Host
4	16	0.0010	0.0808	0.0245	0.0217
5	32	0.0016	0.0654	0.0178	0.0221
6	64	0.0036	0.0834	0.0290	0.0210
7	128	0.0098	0.1246	0.0306	0.0229
8	256	0.0312	0.2029	0.0310	0.0220
9	512	0.1126	0.3500	0.0321	0.0233
10	1024	0.4365	0.6570	0.0361	0.0232
11	2048	1.6847	1.2731	0.0416	0.0214
12	4096	6.6512	2.5056	0.0528	0.0226
13	8192	26.5110	4.9678	0.0777	0.0229
14	16384	105.6375	10.4084	0.1268	0.0231
15	32768	423.4157	23.8199	0.0226	0.1565
16	65536	1704.7605	73.3330	0.2825	0.0226
17	131072	6838.7183	274.7251	0.5295	0.0236

Speed Improvement: Times listed in green represent speed improvement over the host code.
Changes to the Code: Nothing substantial. Defined the max thread count for the computation and also created a shared memory address for the blocks within the grid.
GPU Advantage: The thread and times listed in red represent the thread count when GPU parallelism surpasses the CPU in lower runtime and complexity.