The initial setup of the shared histogram per block gave a speedup of 2.8. This version has bank conflicts. I found a blog that claimed they got maximum efficiency when there were only two blocks running on each sm. I found that number by using a block of code

cudaDeviceProp prop;

checkCudaErrors( cudaGetDeviceProperties( &prop, 0 ) );

int blocks = prop.multiProcessorCount;

This number of multiProcessorCount times two gives seemingly the best number of blocks. Which then took the runtime from about 2.5 down to 1.17ms. I wonder if a different sm break down would provide optimal efficiency on a less optimal histogram size.