1. Tile width used in kernel implementation = 16, so number of threads is 16*16 = **256**

   Maximum number of threads that can be simultaneously scheduled for execution on GeForce GTX 280 GPU, which contains 30 streaming multiprocessors depends on number of blocks that can be used. This can be calculated by considering 4 possible limitations

   1. Based on maximum number of threads:

      Maximum number of threads per blocks = **1024**

      Number of blocks that can be configured for tile length of 16 = **1024/256 = 4** **….. (1)**

   2. Based on total amount of shared memory available:

      Total amount of shared memory available in GTX 280 GPU = 16KB

      Size of shared memory utilized for current kernal implementation with tile size of 16 = **2136 bytes**

      Therefore, Number of blocks that can be configured = **16384/2136 = 7** **…..(2)**

   3. Based on number of registers available:

      Registers used by each threads = **12**

      Total number of registers available per block: **16384**

      Therefore, Number of blocks that can be configured = **16384/ (12*256) = 5** **…...(3)**

   4. Based on maximum number of blocks that can be assigned to each SM

      Maximum number of blocks that can be assigned to each SM = **8** **….. (4)**

   So by considering all the 4 limitations, we can come to the conclusion that the main limiting factor is going to be the maximum number of threads per block, which is **= 4**

   So Maximum number of threads that can be simultaneously scheduled for execution =

   = **blocks/SM * threads/block * SM/GPU**

   = **4 blocks/SM * 256 threads/block * 30 SM/GPU = <u>30720 threads</u>**