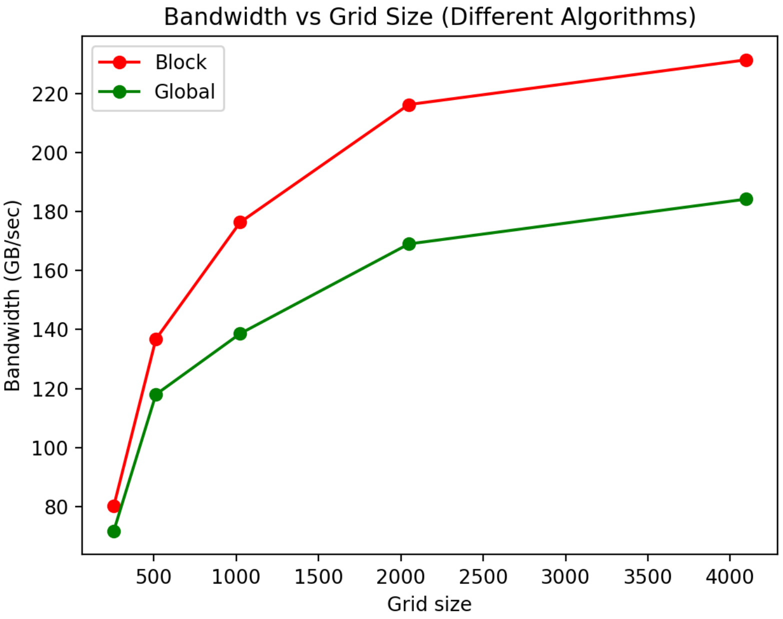
**CME213: Parallel Computing using MPI, OpenMP and CUDA**

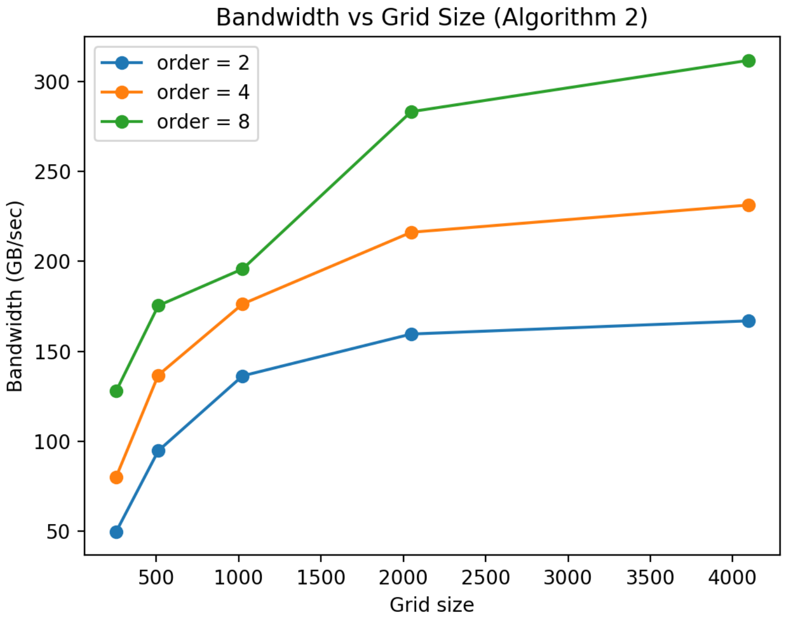
**Homework 3: Solving 2D PDE with CUDA** **Chi Zhang**

Stanford University SUID: 06116342

Spring 2018 Date: May 9, 2018

# Question 3





## Question 4

Among the first two algorithms using global memory, algorithm 2 performs much better than algorithm 1. As mentioned in class, square blocks typically perform better than rectangular ones because its ratio of flops/word is higher. Concretely, algorithm 1 uses 192 threads per block with shape of (32, 6) while algorithm 2 uses 256 threads per block with shape of (32, 8). However, since each thread in the block loops multiple times until the whole (32, 32) chunk has been processed, the effective block size in algorithm 2 becomes (32, 32), which is a square. Furthermore, there may also exist small performance improvement when setting up warps less times in algorithm 2.

For a $n \times n$ square block, memory traffic is $(2n^2 + 4n)$ and flops is $10n^2$. Denote the intensity as $I(n)$:

$$n = 2,\ I(2) = 2.5$$

$$n = 4,\ I(4) = 3.33$$

$$n = 8,\ I(8) = 4$$

Therefore, in this problem, the best performance is achieved when order is 8. In general, it can be found that the maximum asymptotic intensity is 5 flops/words, and kernel with higher-order compact stencils will have better performance.