Tian Tang (A20284409)

CUDA programming design: I noticed that the mean and deviation of each column in matrix are relatively independent from another column. So I used one thread to deal with one column. I had the blocks with 32*1*1 threads and each grid with N/32 blocks inside. I experimented with different matrix size and compare the elapsed time with serial codes. As the results shown below, the elapsed time of serial codes increases with the growing of matrix size. However, the calculating time of CUDA program barely increases and stays relatively stable. When the size of matrix is small, speedup is less than 1 due to the overhead. But it grows rapidly and becomes more than 1 after some threshold (around 5000).

Matrix Size	Serial/CUDA	Elapsed Time	Speedup
100	S	0.474	
100	C	586.098	0.001
500	S	0.591	
500	C	564.048	0.001
1000	S	13.893	
1000	C	573.223	0.024
2000	S	59.074	
2000	C	560.315	0.105
5000	S	416.931	
5000	C	556.24	0.750
6000	S	655.385	
6000	C	562.054	1.166
7000	S	878.504	
7000	C	551.115	1.594
8000	S	1308.48	
8000	C	556.465	2.351