



The reason that the run time did not show significant decrease after reaching a certain number of threads is because when the thread number went high, threads and data would be spread in different cores and different banks. Therefore, hops of memory access would happen which were costly and prevented the performance from the further improvement