I was finally able to implement all 6 algorithms in CUDA. Histogram was successful, but I stuck with gpu reduction. I tried to debug several times, but none of them work. Then I realized that the problems are that when there’s extra threads compared to limited array size, they should read different values within different operations. For example, for sum, the empty threads should read 0; for product, it should read 1; for min, it should read INT\_MAX; for max, it should read INT\_MIN. I set the empty threads reads in identical values for all four operations, that’s why I failed to debug every time.