Hello everyone, next, Chuzihao, Wangfu and I will introduce the design and the results of our program for this project “Hadoop Map-Reduce from scratch”. Our presentation contains four parts. In the first part, I will briefly introduce the design of our program. Then, Chuzihao will run the codes on his computer and show the details. At last, Wangfu will give the comparison of the performance of two methods and possible improvements of our program.

The purpose of the project is to compare the performance of the trivial method and the Map-Reduce method for word counting process.

For the trivial method, we do the splitting, reducing work on one processor and apply a quick sort without recurrence to obtain the final result. For the Map-Reduce method, we use docker to simulate the slaves. We created three containers as the three slaves. All of them are connected in one virtual network in order that they can transfer files to each other.

Here is the system diagram of our program design. The first step is that the master reads all the files line by line and write them into three files Sx1, Sx2, and Sx3, where each file contains a third of the original file. The transmission of the three files is done by copying them to correspondent slaves. Then the three slaves separate the file received to three child files and send to itself and the other two slaves. The rule is to send the words whose first letter’s ASCII code is smaller than capital letter R to slave1, the words whose first letter’s ASCII code is between capital letter R and little i to slave2, while the other words to slave3. Here the transmission is done by using scp service. Then each slave does the shuffling and reducing work and send back their results to the master. The master use these results to do a quick sort to obtain the final result.