Abstract

Designing Hadoop cluster on a virtual network is a great way to increase the number of nodes that serve as tasktracker that has responsibility to perform data processing on a cluster. That is the way to improve mapreduce performance. The basic theory is the more nodes serves as tasktracer, the faster the process running.

On this final project Hadoop Cluster will be designed with Hadoop Distributed File System (HDFS) as the file system and MapReduce as the programming model for data processing. The addition of the node will be conducted on three physical machines and maximum four virtual machine on each node. Virtualization approach used in cluster is OS-level-virtualization useing OpenVZ. Clusters running on virtual environments will be tested by looking at the throughput of the HDFS and the execution time of sort of testing. The results will be compared with MapReduce that run only on physical nodes. From the testing results and performace analysis, it may be concluded that the cluster with physical scale is better than virtual scale. Addition in every three virtual node raises the average of execution time of 7.03% (3.69 seconds).

Keywords: Hadoop, cluster, *mapreduce*, distribution, virtualization.