World Scientific Research Journal Volume 5 Issue 7, 2019
ISSN: 2472-3703 DOI: 10.6911/WSRJ.201907_5(7).0020

The Implementation of Improving the Efficiency of Network Data Transmission Based on Deduplication

Youping Dong[1, a]

[1]School of Information Science and Electrical Engineering, Shandong Jiaotong University, Jinan, China

[a]156742249@qq.com

Abstract

In this paper, Deduplication is discussed as a green storage technology to improve the efficiency of network data transmission. Firstly, this paper introduces the concept of deduplication and what kinds of normal fields it can be used. Secondly, it explains the working principle of deduplication. Lastly, the method is implemented to improve the efficiency of network data transmission by software instead of improving the hardware system as before. It analyzes the process how to increase speed of network data transmission by using deduplication.

Keywords

Deduplication, Network data transmission, Bandwidth, Hash table.

1. INTRODUCTION

Deduplication is a technique for keeping only one copy of the data by deleting the repeating ones. It can largely reduce the requirements of physical storage space which is becoming more and more valuable in now days. Deduplication can bring a lot of benefits as it is well known that big data is getting more and more popular. For example, improving the efficiency of the storage system, reducing the demand of network bandwidth during transmission. It is a green storage technology to save costs and decrease consuming effectively.

2. WORKING PRINCIPLE

According to the granularity, deduplication is divided into two levels that are based on file and data block. Deduplication of the file level is single instance storage while the one of data block level takes the smaller grain size which can get the size from 4 to 24kB.The level of data block is higher in efficiency to delete repeated data. At present it is popular for the deduplication products to be on the level of data block. It is the way to divide the file into data blocks, which are fixed length or not, and then calculate their indexes by the Hash algorithm. At present, the popular Hash algorithms are MD5 and ShAI. Sometimes two or more kinds of algorithms are used to make the probability of data collision less enough to be ignored. The data blocks with the same indexes are considered the same blocks which should be remained only one copy. So a block corresponds an index. The index table will be read before the file. According to the index, the corresponding data block is read from the storage system to restore the copy of the file.

So far deduplication is used mainly in data backup. It is the right way to delete the repeated data which are made by many times backup. The data deletion rate is reported from 20:1 to 500:1. In addition, deduplication is also implemented in other fields, such as online data storage, file system, network data transmission and so on.

3. IMPLEMENTATION

This paper introduces how to improve the speed of network data transmission by using the technology of deduplication. So far increasing the speed of network data transmission can only come about through improving the hardware system which can't change the barrier among the operators. But the method in this paper can solve this problem by software.

This paper describes the method which improves the speed of network data transmission on the fixed bandwidth. Figure 1 shows the structure of network data transmission based on deduplication. There is a client named C will download data from a server named S, but the line between S and C is low-bandwidth and high delay and C will download the same or some of the same data many times. Because of the large data in network traffic, normally C can only wait. However, the method here sets up a proxy server on the both sides of C and S. They are named Pc and Ps. If it is the first time for C to download the data from S. While the data is transferred from S to C for the first time, the data is divided into blocks and indexes are calculated to establish the Hash table. At the same time, the Hash table is also created and data blocks saved on Pc. Note: the Hash tables are the same on both Ps and Pc. If C downloads the data from S for the second time or more later. Ps divides the data into blocks at first, and then search the indexes of the blocks in the table. If the indexes of the data blocks are not found in the table, then the blocks are sent to Pc and saved on Pc. The corresponding indexes are added to the tables on both Ps and Pc. The data blocks are sent to C; If the indexes of the data blocks are found in the table on Ps, then Ps notifies the indexes of the blocks to Pc. Pc sends the corresponding blocks to C according to the indexes. In this case, the remote data transmission from S to C is changed to the local transmission from Pc to C. So the method greatly reduced the time and bandwidth requirements. In the process, because only the deduplicated data blocks are saved on Pc the data transmission from S to C sometimes becomes local transmission. Only if the data never transferred before is belong to remote transmission.

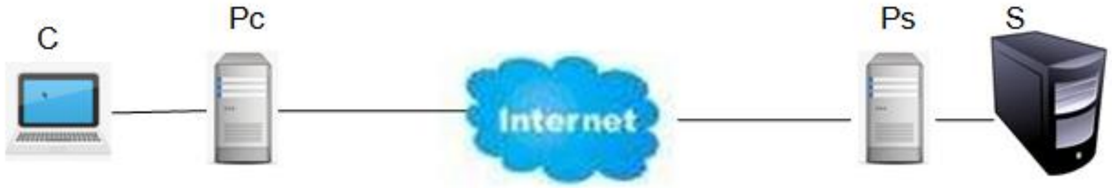

Figure 1. Structure of network data transmission based on deduplication

Deduplication in this paper is implemented through the technology of setting the size of the data block, generating the indexes, searching data. The size of the data block can be set according to the actual situation. The indexes can be created by Hash function adopting MD5 which can achieve the goal of different data block with different index. Searching data blocks uses Hash search which is well known as its great speed of searching. So it is a good choice in the situation that the large network data transmission needs great requirement of search performance.

Based on deduplication, the data coming from the same source are deduplicated on the source side and saved on the target side. The part of deduplicated data or only the indexes of the reduplicated data are transferred. So it can save network bandwidth. Although it occupies a lot of space to save the deduplicated data on Pc, it is still a good method to improve the efficiency of network data transmission by sacrificing spatial complexity.

REFERENCES

[1] Information on https://blog.csdn.net/qq_31909617/article/details/78396679

[2] Chen Yue: Data Structure (Higher Education Press, China 2016), p.166-189 (In Chinese).

[3] Information on https://github.com/adamierymenko/kissdb

[4] Information on https://www.openssl.org/docs/man1.1.0/man3/MD5_Init.html