Deduplication for Chunk-Based File Backup

Yu Yu Maw, Khin Mar Soe
University of Computer Studies, Yangon
yuyumaw.89@gmail.com

Abstract

Data deduplication has become a popular technology for reducing the amount of storage space necessary for backup and archival data. As the amount of data growth all over the world, many organization need to reduce amount of data safely. To protect size of data enormously growth, deduplication become a solution to solve this problem. Data deduplication is found in many forms. Data deduplication reduces the data that duplicate within a file or among other files. Virtual tape libraries, archive storage, disk storage systems, and applications such as email systems, content managers, backup systems and more, are examples of where data deduplication can be applied. It can achieve more storage space although there are multiple files. So, data deduplication becomes essential and critical component of backup systems.

In this thesis, deduplication for chunk-based file backup is implemented using Content-Defined Chunking Algorithm and Secure Hash Algorithm. In this thesis, deduplication works with four steps: (1) Chunking (2) Fingerprinting (3) Index lookup (4) Writing. ***Content-Defined Chunking Algorithm*** *chunks input file stream to generate chunks.* ***Secure Hash Algorithm*** *is use to generate hash key (fingerprint) and to compare new chunks and old chunks in order to get unique chunks in the file. In such system, we focus on Microsoft Word files. In normal, deduplication takes a long time to complete and use a lot of CPU cycles in the process of deduplicating data, possibly introducing performance issues on production machines. Instant of deduplication make over entire file, this thesis dedupes a file by partition three parts in order to reduce processing time.*