Researchers conduct survey on deduplication systems

May 1 2024, by David Bradley

Credit: Pixabay/CC0 Public Domain

A review published in the *International Journal of Grid and Utility*

Computing has investigated ways in which the increasing problem of duplicate data in computer storage systems might be addressed. Solutions to this problem could improve storage efficiency, system performance, and reduce the overall demand on resources.

Amdewar Godavari and Chapram Sudhakar of the department of Computer Science and Engineering at the National Institute of Technology Warangal in Warangal, Telangana, India explain how the advent of the Internet of Things (IoT) and the emergence of big data in science, engineering, medical, and many other areas has led to a massive increase in computer storage demand.

Some researchers have suggested that by 2025, the amount of stored data will amount to around 175 zettabytes (175 trillion terabytes). Other research has provided estimates of duplication in this data and suggests that around three-quarters, 75%, is wholly redundant. This redundancy leads to inefficient storage utilization and decreased performance in storage systems. Identifying the duplicate content that might be removed from a system is not a simple matter.

To address this challenge, the researchers point out that there are two general approaches. The first is data compression, which will compare files and crush file sizes based in the identification of duplicates. Full-on data deduplication, however, can compute a unique "hash value" for much larger blocks of data, compares those hashes to find blocks containing identical data and so flag them for removal as appropriate. This latter approach could be used to reduce the amount of down-time or latency that would otherwise impinge on performance and access.

The team suggests that various chunking algorithms and machine learning-based techniques might be used to identify redundant blocks of data. Their tests show that variable-sized chunking offers better deduplication ratios compared to fixed-sized chunking, although this

approach is slower. The algorithmic approach, however, could allow redundancy categorization to use machine learning to improve efficiency still further.

More information: Amdewar Godavari et al, A survey on deduplication systems, *International Journal of Grid and Utility Computing* (2024). DOI: 10.1504/IJGUC.2024.137902