Fragmentation Study for Deduplication in cache Backup Storage

M.Sakthivel¹, Karnajoy Santal²,Bhaskar Rao K³

¹Veltech Hightech Engineering college, Department of Computer Applications,
Avadi, Chennai 600062, India
vel.sakthi@gmail.com

²Master of computer application, VelTech Hightech Engg. college,
171, sri bharathi nagar, avadi, chennai-600062, india
karnajoy2012@gmail.com

³Master of computer application, VelTech Hightech Engg. college,
Reshmi illam, no.2/228/a, TSP camp road, veerapuram, avadi, chennai-600055, india
bhaskar0498@gmail.com

Abstract:

In backup environments field deduplication yields major advantages. Deduplication is process of automatic elimination of duplicate data in storage system and it is most effective technique to reduce storage costs. De duplication effects predictably in data fragmentation, because logically continuous data is spread across many disk locations. Fragmentation mainly caused by duplicates from previous backups of the same back upset, since such duplicates are frequent due to repeated full backups containing a lot of data which is not changed. Systems with in-line deduplicate intends to detects duplicates during writing and avoids storing them, such fragmentation causes data from the latest backup being scattered across older backups. This survey focused on various techniques to detect inline deduplication. As per literature, need to develop a focused on deduplication reduce the time and storage space. Proposed novel method to avoid the reduction in restores performance without reducing write performance and without affecting deduplication effectiveness.

Keywords: Chunking, De duplication, Fragmentation, Recovery.

1. Introduction:

In advanced technology’s world with variety of application leads to growth of stored digital information, duplicate data is receiving increased attention. From last few years archival and backup systems used automatic removal technique to remove duplicate data and recently have become characteristic of several storage appliances. Since last few decades deduplication gained great popularity in field of storage. The effectiveness of such method in decreasing both time required completing backups and storage space is necessary to save them. Aside from its write performance, read performance of the deduplication storage has been attaining in importance with a wide range of its utilizations. Deduplication is defined as a process of automatically eliminating coarse-grained and unrelated duplicate data. Main objective of deduplication is to eliminate both interfile and interfile redundancy over large datasets, stored at different times by uncoordinated users. Traditionally, performance of a deduplication system is measured by the data deduplication ratio, maximal write performance and restores bandwidth. The read bandwidth is usually very good for an initial backup saved to an empty system, but depreciates for subsequent backups. The occurrence of this issue is due to data fragmentation caused by
in-lined duplication which results in data reasonably belonging
to are cent backup distributed through multiple older backups.
Fragmentation mainly caused by duplicates from previous
backups of the same backup set, since such duplicates are
frequent due to repeated full backups containing a lot of data
which is not changed. The remaining paper can be summed in
next section studies some earlier developed methods for in-line
deduplication.

2. Related Work

The proposed scheme provides two-fold approach, first, a
novel indicator for dedupe scheme called cache-aware Chunk
Fragmentation Level (CFL) monitor and second selective
duplication for improvement read performance. The CFL is
consists of two parameters: optimal chunk fragmentation and
cache-aware current chunk fragmentation. Whenever the
current CFL becomes worse than the demanded one, then
selective duplication technique is activated to enhance read
performance. Proposed scheme assures demanded read
performance of each data stream while completing its write
performance at practical level and also guaranteed a target
system recovery Time. Major drawback of selective
duplication is that it requires extra memory space called in
memory temp container. To improve the efficiency of
typical compressors in modern storage systems, author
proposed Migratory Compression (MC), a coarse-grained data
transformation. Relocation of similar data chunks gathers into
one unit, to improve compression factors. After decompression
migrated chunks return to their previous locations. While
compressing single files, MC paired with a typical compressor
such as grip provides a clear improvement. Proposed scheme
improves compression effectiveness by 11% to 105%,
compared to typical compressors. Deduplication process is
use delaminating duplicates in data, thus improving the
effective capacity of storage systems. Single-node raw capacity
is still mostly limited to tens or a few hundreds of terabytes,
forcing users to resort to complex. In [4], author proposed new
mechanisms called progressive sampled indexing grouped
mark and sweep, to address dedupe challenges and also to
improve single node scalability. Progressive sampled indexing
removes scalability limitations by using indexing technique.
Advantages of proposed scheme are, improves scalability,
provide good deduplication efficiency and improvement in
throughput. The proposed three fold approaches, first they
discuss sanitization requirements in the context of deduplicated
storage second implemented a memory efficient technique for
managing data based on perfect hashing, third they design
sanitizing deduplicated storage for EMC data domain.
Proposed approach minimizes memory and I/O requirements.
Perfect hashing requires a static fingerprint space, which
conflicts with proposed scheme desire to support host writes
during sanitization Data deduplication has recently gain
importance in most secondary storage and even in some
primary storage for the storage purpose. A read performance of
the deduplication storage has been gaining great significance.
In [6], introduced called Chunk Fragmentation Level (CFL) as
an Indicator for read performance. Proposed approach is very
effective to indicate the read performance of a data stream with
deduplication storage. Proposed the context based rewriting
algorithm which rewrites selected few duplicates proposed
algorithm recovers restore speed of the latest backups, while
resulting in fragmentation increased for older backups, which
are rarely used for restore. Main drawback proposed algorithm
is that it reduces write performance a little bit. In this paper [8],
author proposed an inline deduplication solution for primary
workloads. Proposed algorithm is based on two key insights
from real world workload that is spatial locality exists in
duplicated primary data and second temporal locality occurs in
the entrance forms of duplicated data. Proposed algorithm
minimizing extra Input Outputs and seeks.

3. Client Server Architecture

Figure 1: Client-Server Backup Storage
In client server architecture interaction client wants to take backup of storage system on the server side. So clients send request to the server. When server approves client request for Backup then connection get formed between client and server. Server acquires backup of client storage and proceed for storage purpose. In this process we should assure that performance measures and restore should be good during client request for restoration. Server should be available and very quick to give access to that backup whenever client or user wants to access data from backup storage on server side. At the server side, server takes backup and proceeds for fragmentation. File level fragmentation technique fragments the backup set into small files so that it will be easy to analysis and manage for storage and restore purpose. Those fragmented files get divided into segments and get stored into containers. Segment index get stored in index module for restore purpose. Duplicate data index value do not get stored, if same data value get detected during hashing that time same data value get referenced to already existing data value index. Deduplication technique gives following benefits:

i) Read data from files
ii) Stored into segments
iii) Calculate index values
iv) Stores those indices.

Throughput and I/O performance should be good during restoration process. So it is need to develop new techniques which are more capable for changing technology and faster data process with large data set [4].

4. Comparison

Assuring demanded read performance of data deduplication storage with backup data sets, in this paper they proposed scheme provides two-fold approach, first, a novel indicator for dedupe scheme called cache-aware Chunk Level (CFL) monitor and second selective duplication for improvement read performance. Proposed scheme assures demanded read performance of each data stream while completing its write performance at a practical level and also certain a target system recovery time. Major drawback of selective duplication is that it requires extra memory space called in-memory temp container Migratory Compression: Coarse-grained Data Reordering to Improve Compressibility in this paper author improved the efficiency of traditional compressors in modern storage systems author proposed Migratory Compression (MC), a coarse-grained data transformation. Proposed scheme improves compression effectiveness by 11% to 105%, compared to traditional compressors. Disadvantage is overhead to perform Building a high performance deduplication system in this paper author proposed new mechanisms called progressive

5. CLOUD V/S TRADITIONAL BACKUP

Table 1. Cloud V/S Traditional Backup [4]

<table>
<thead>
<tr>
<th>Constraint</th>
<th>Cloud Backup</th>
<th>Tradition Backup</th>
</tr>
</thead>
<tbody>
<tr>
<td>Users</td>
<td>Single</td>
<td>Enterprise</td>
</tr>
<tr>
<td>Backup sizes</td>
<td>Minimum</td>
<td>Maximum</td>
</tr>
<tr>
<td>Bandwidth Address</td>
<td>Low Bandwidth</td>
<td>High Bandwidth</td>
</tr>
<tr>
<td>Storage Type</td>
<td>Cloud Server</td>
<td>Data Server</td>
</tr>
<tr>
<td>Restore</td>
<td>Segment</td>
<td>Container</td>
</tr>
<tr>
<td>Access</td>
<td>More Flexible</td>
<td>Less Flexible</td>
</tr>
<tr>
<td>Examples of System</td>
<td>Backup</td>
<td>Symantec</td>
</tr>
</tbody>
</table>

Sampled indexing and grouped mark and sweep, to address dedupe challenges and also to improve single-node scalability. Advantages of proposed scheme are, improves scalability, provide good deduplication efficiency and improvement in throughput. The indexing metadata size grows linearly with the capacity of the system [4]. Memory efficient sanitization of a deduplicated storage system in this paper author proposed three fold approaches, first they discuss sanitization requirements in the context of deduplicated storage, second implemented a memory storage efficient technique for managing data based on perfect hashing, third they design sanitizing deduplicated storage for EMC data domain. Proposed approach minimizes...
memory and I/O requirements. Using perfect hashes requires a static fingerprint space, which conflicts with proposed scheme desire to support host writes during sanitization [5]. Reducing impact of data fragmentation caused by in-lined duplication author proposed the context based rewriting algorithm which rewrites selected few duplicates. Proposed algorithm recovers restore speed of the modern backups, while resulting in fragmentation increased for older backups, which are rarely used for restore. Main drawback proposed algorithm is that it reduces write performance a little [7].

Table 1 displays the difference amid cloud backup and established backup. In established backup, chunks are transferred to the client afterward being acquired from the containers in the data servers in reinstate process. If a data server is running countless jobs at a period, reading data from disk will come to be a block in established backup. As slender cloud arrangement services interface of reading a finished file (i.e., segment), data is elucidate in the client afterward the segments are transferred from the cloud in cloud backup. Overall, the speed of reading data from disk is far faster than that of elucidating data across Expansive Expanse Network. Thus, the bottleneck of reinstate procedure in cloud backup is the segment advancing procedure [4][9].

6. Conclusion

This paper presented an inclusive survey on the in-lined duplication reduction. The main features, the advantages and disadvantages of each detection technique are described In backup environments field called deduplication yields major advantages and also results certainly in data fragmentation, because rationally constant data is spread across many disk locations. As per survey, there is strong need of a fast, robust and innovative method to detect in-line deduplication to reduce time needed to perform backups and storage space required to save them.

7. References