3Gen Data Deduplication Technical Discussion – White Paper

3Gen Data Systems

3Gen Data Deduplication Technical Discussion – White Paper

Table of Contents

3Gen Data Systems

3Gen Data Deduplication Technical Discussion – White Paper

Introduction

In the deduplication process, unique chunks of data, or byte patterns, are identified and stored during a process of analysis. As the analysis continues, other chunks are compared to the stored copy and whenever a match occurs, the redundant chunk is replaced with a small reference that points to the stored chunk. Given that the same byte pattern may occur dozens, hundreds, or even thousands of times (the match frequency is a factor of the chunk size), the amount of data that must be stored or transferred can be greatly reduced.

The process where the deduplication hash calculations are created on the target device as the data enters the device in real time, may occur "in-line", as data is flowing, or "post-process" after it has been written. If the device identifies a block that it already stored on the system, rather than storing the new block, it references to the existing block. The benefit of in-line deduplication over post-process deduplication is that it requires less storage as data is not duplicated. On the negative side, it is frequently argued that because hash calculations and lookups takes so long, it can mean that the data ingestion can be slower thereby reducing the backup throughput of the device. However, this equation changed with the advent of a new generation of higher performance multi-core processors and mass market versions of high performance storage media, including solid state and high-rpm SAS drives. That combination made it possible for 3Gen to design deduplication systems that could provide postprocessing levels of performance using a more traditional inline data flow. 3Gen went through a year-long development and testing by using an approach designed specifically to take advantage of the new hardware platforms. All of 3Gen's current V9000Pro Storage platform employ the full inline data flow which provides equal or better performance than our competitor's data deduplication solutions.

Implementation

The purpose of data deduplication is to increase the amount of information that can be stored on disk arrays and to increase the effective amount of data that can be transmitted over networks. When it is based on variable-length data segments, data deduplication has the capability of providing greater granularity than single-instance store technologies that identify and eliminate the need to store repeated instances of identical whole files. 3Gen Data deduplication operates by segmenting a dataset -this is normally a stream of data— into blocks and writing those blocks to a disk target. To identify blocks in a transmitted stream, the data deduplication engine creates a digital signature for each data segment

and an index of the signatures for a given repository. The index, which can be recreated from the stored data segments, provides the reference list to determine whether blocks already exist in a repository. The index is used to determine which data segments need to be stored and also which need to be copied during a replication operation. When data deduplication software sees a block it has processed before, instead of storing the block again, it inserts a pointer to the original block in the dataset's metadata. If the same block shows up multiple times, multiple pointers to it are generated. Variable-length data deduplication technology stores multiple sets of discrete metadata images, each of which represents a different dataset but all of which reference blocks contained in a common storage pool (Figure 1) . 3Gen data deduplication architecture is based on years of proven research and commercial implementation of IBM Snapshot. This technology provides our customers many years of worry free data protection.

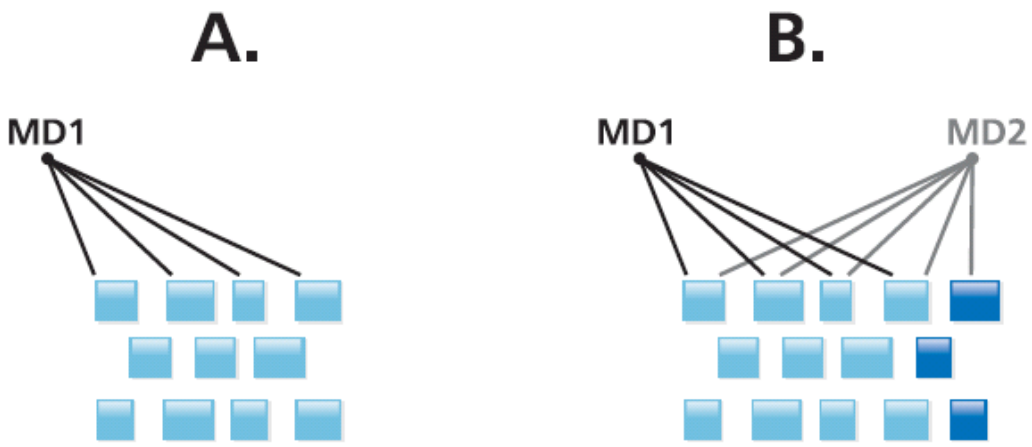


Figure 1. Data Deduplication Methodology

Fixed length blocks vs Variable length data segments

It is possible to look for repeated blocks in transmitted data using fixed-length block divisions, and that approach is currently being used by many storage suppliers to include deduplication as a feature of the software. Fixed block systems are used most often when general purpose hardware is carrying out deduplication because less compute power is required. The tradeoff, however, is that the fixed block approach achieves substantially less effective reduction than a variable-block approach. The reason is that the primary opportunity for data reduction in a backup environment is in finding duplicate blocks in two transmitted data sets that are made up mostly—but not

completely—of the same segments. If we divide a backup data stream into fixed-length blocks, any change to one part of the dataset normally creates changes in all the downstream blocks the next time the data set is transmitted. Therefore, two data sets with a small amount of difference are likely to have very few identical blocks

Figure 2 applying fixed block lengths to a data sequence: the upper line shows the original block division—the lower line shows the blocks after making a single change to Block A (an insertion). In spite of the fact that the shaded sequence of information is identical in the upper and lower lines, all of the blocks have changed content and no duplication is detected. If we stored both sequences, we would have 8 unique blocks.

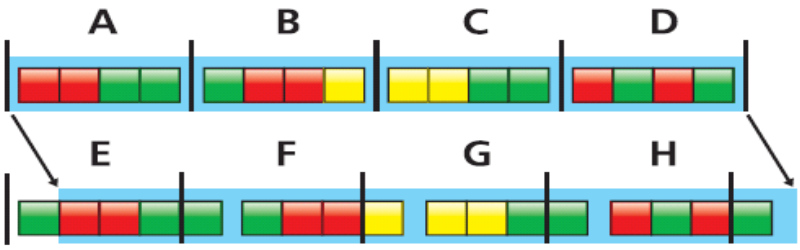


Fig. 2 = Fixed length blocks

Instead of fixed blocks, 3Gen's deduplication technology divides the data stream into variable length data segments using a data-dependent methodology that can find the same block boundaries in different locations and contexts. This block-creation process allows the boundaries to "float" within the data stream so that changes in one part of the dataset have little or no impact on the boundaries in other locations of the dataset. Through this method, duplicate data segments can be found at different locations inside a file, inside different files, inside files created by different applications, and inside files created at different times.

Figure 3 applying variable-length segmentation to a data sequence: In this case, Block A changes when the new data is added (it is now E), but none of the other blocks is affected. Blocks B, C, and D are all recognized as identical to the same blocks in the first line. If we stored both sequences, we would have only 5 unique blocks.

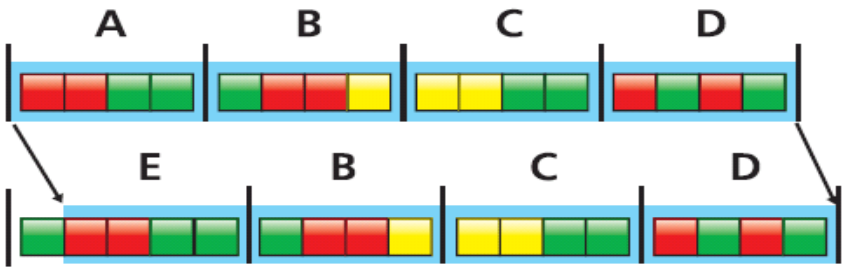
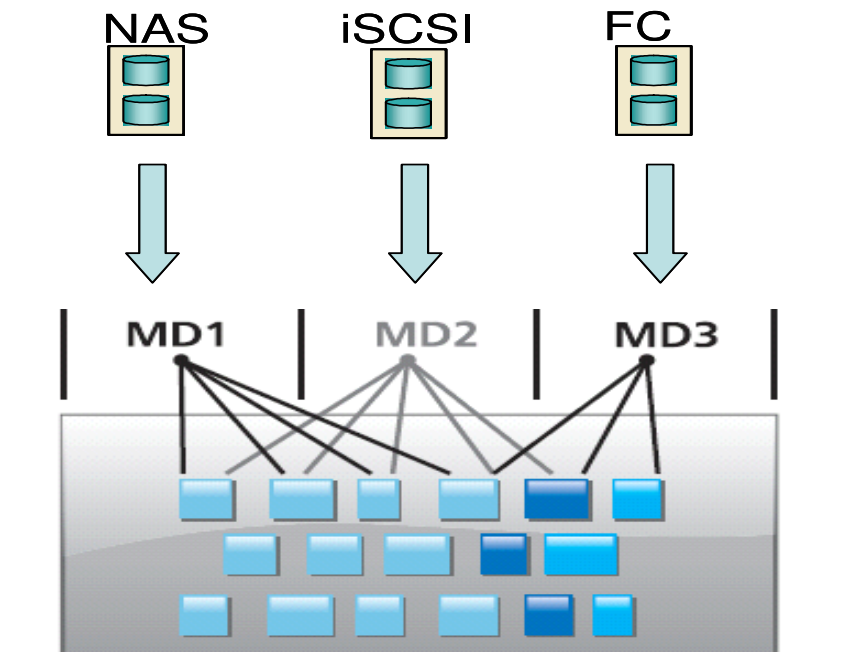


Fig 3 = Variable-length segmentation

Sharing a Common Deduplication Block Pool

Data deduplication systems gain the most leverage when they allow multiple sources and multiple system presentations to write data to a common, deduplicated storage pool. 3Gen V9000Pro provides access to a common deduplication storage pool through multiple presentations that may include a combination of NAS (CIFS or NFS), iSCSI and FC volumes. Because all the presentations access a common storage pool, redundant data segments are eliminated across all the datasets being written to the V9000Pro Storage Systems. In practical terms, this means that 3Gen V9000Pro will recognize and deduplicate blocks that come from different sources and through different interfaces—for example, the same data segments on a print and file server backed up via NAS and on an email server backed up via FC volume.



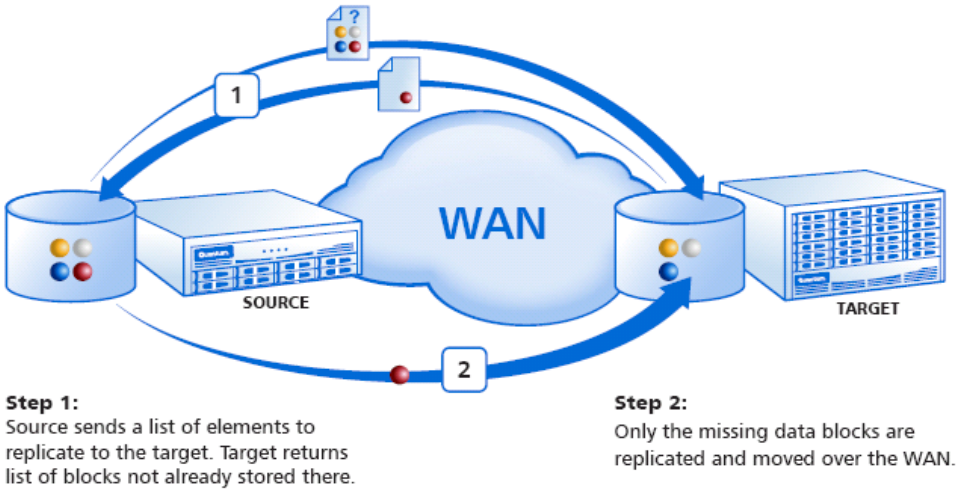
3Gen V9000Pro Storage

Applying Data Deduplication to Replication

Up to now, our discussion has focused primarily on the storage benefits of deduplication, but the technology provides similar benefits to remote replication by dramatically reducing the bandwidth needed to copy data over networks. The result gives disk backup a practical way to provide WAN-based disaster recovery (DR) protection and to reduce requirements for removable media.

The minimum Disaster Recovery (DR) protection required from every IT organization is ensuring that backup data is safe from site loss or damage. Equipment and applications can be replaced eventually, but digital assets are often irreplaceable. No matter how resilient or redundant a given storage or backup system may be or how many layers of redundancy it might have, when all copies of data are located at a single site and in a single hardware system, they are vulnerable to site-specific damage, including natural disasters, fire, theft, and malicious or accidental equipment damage. Data deduplication technology gives IT departments a new DR option by making inter-site replication over

WANs a practical alternative that can enhance DR preparedness, reduce operating expenses, and decrease the usage of removable media.



Data deduplication makes the process of replicating backup data practical by reducing the bandwidth and cost needed to create and maintain duplicate datasets over networks. At a basic level, deduplication-enabled replication is similar to deduplication-enabled data stores. Once two images of a backup data store are created, all that is required to keep the replica or target identical to the source is the periodic copying and movement of the new data segments added during each backup event, along with its metadata image, or namespace.

Bandwidth and Storage Capacity Savings

The extent of data deduplication depends upon the number of users and the type of data they generate. But beyond a certain number of users, the amount of unique content is capped by the organizations ability to generate unique data.

To establish the benefits of deduplication, we benchmarked some of our key customers based on following parameters –

1. No. of users
2. Avg. Data / PC and avg. daily change
3. Type of data (emails/ documents/ media etc.) and retention period

Customers	No. of Users	Avg. Data/ PC (GB)	Average Daily Changes (%)	Data Type	Retention Period (Days)	Total Storage (TB) before 3Gen Data Deduplication	Total Storage (TB) after 3Gen Data Deduplication
Large Financial Company	2000	20	5	Docs and email	90	60	12
Petroleum Corporation	500	6	2	Emails	30	10	1.2
Consultancy Group	300	10	2	Docs and Emails	90	27	2
Small Web Designer Company	100	45	1	Mostly Media	15	6.8	1.6
District High School	350	80	10	Emails, Doc and Media	90	45	20

Information about 3Gen V9000Pro Data Deduplication Storage Platform

3Gen V9000Pro extends the benefits of data deduplication across the Enterprise and integrates it with replication, and encryption into a complete storage solution for multi-site environments. 3Gen's patented variable-length data deduplication reduces typical disk requirements by 90% or more and makes WAN-based replication a practical DR tool. The result is fast, reliable backup and restore, reduced media usage, reduced power and cooling requirements, and lower overall data protection and retention costs. The V9000Pro provides disk storage solutions with deduplication and replication for use in a wide range of IT environments. All V9000Pro systems are based on a common foundation and can be linked through replication to provide a multi-site protection strategy

For more information about 3Gen V9000Pro Data De-duplication systems, please contact our regional sales offices and visit 3Gen website, www.3gendata.com