SNIA
Data Management Forum

Understanding data deduplication ratios

June 2008

Mike Dutch
Data Management Forum
Data Deduplication & Space Reduction SIG Co-Chair
EMC Senior Technologist

Table of Contents

List of Tables

List of Figures

Optimizing storage capacity

Data deduplication and other methods of reducing storage consumption play a vital role in affordably managing today's explosive growth of data. Optimizing the use of storage is part of a broader strategy to provide an efficient information infrastructure that is responsive to dynamic business requirements. This paper will explore the significance of deduplication ratios related to specific capacity optimization techniques within the context of information lifecycle management.

The benefits of optimizing storage capacity span cost savings, risk reduction, and process improvement. Capital expenditures on networked storage equipment and floor space can be reduced or deferred. Ongoing operating expenses for power, cooling, and labor can also be reduced because there is less equipment to operate and manage. Increasing the efficiency and effectiveness of their storage environments helps companies remove constraints on data growth, improve their service levels, and better leverage the increasing quantity and variety of data to improve their competitiveness.

The impact on storage utilization

Capacity optimization refers to any method which reduces the consumption of space required to store a set of data including compression, single instance storage, data deduplication, thin provisioning, copy on write, and pointer remapping. Each of these techniques has a valuable role to play in improving data storage efficiency and may be used in concert to achieve the right balance of resource utilization for specific situations. Before defining each of these terms, it is interesting to note the impact of capacity optimization on storage utilization from two perspectives. The "fit more in a bag" view ascribes goodness to using an existing storage system to hold more data online longer. The "use a smaller bag" view notes that capacity optimization may provide opportunities to reduce or postpone expenditures.

Multiple technologies

The terminology used by the press, analysts, and vendors to refer to capacity optimization techniques often serves to obscure an objective evaluation of customer choices.   Each term has been used as an umbrella term for one or more space reduction techniques. A practical definition of several terms is provided here to convey the essential similarities and differences of each technique and provide a basis for understanding the subsequent discussion of space reduction ratios and percentages.

***Single instance storage* (SIS)** is the replacement of duplicate files or objects with references to a shared copy. An object is a set of data meaningful to an application such as virtual machine images, virtual tape cartridges, disk volumes, email messages, database records, and object-based storage device objects.

Data deduplication is the process of examining a data set or byte stream at the sub-file level and storing and/or sending only unique data. There are many different ways to perform this process but the key factor distinguishing data deduplication from SIS is that data is shared at a sub-file level.

Compression is the encoding of data to reduce its storage requirement. Lossless data compression methods allow the exact original data to be reconstructed from the compressed data while lossy data compression methods permanently discard some of the original data.

Lossless methods are used to compress executable programs and text-based data (such as source code and XML) and where loss of fidelity is not considered acceptable (such as for medical imagery and high definition audio). Example file formats that use lossless data compression include ZIP, GIF, PNG, FLAC, and Dolby TrueHD. Audio and visual information is usually stored in file formats that use lossy compression because superior space savings can be achieved with minor if any loss in perceived quality. Example file formats that use lossy data compression include formats for music (AAC, MP3, Vorbis, WMA), speech (CELP, Speex), images (JPEG), and video (AVC/H.264, Theora, WMV).

Copy on write* and *pointer remapping techniques are used to create changed block point in time copies. Unchanged blocks are shared between the source copy and its snapshots in a manner reminiscent of data deduplication. However, data deduplication has not traditionally been used to refer to the process of sharing common storage between a source copy and its snapshots.

Thin provisioning is the transparent allocation of physical storage space for data when it is written ("just in time") rather than in advance of anticipated consumption. Space savings stems from avoiding media pre-allocation although additional space saving techniques can also be applied.

Complementary technologies

Capacity optimization technologies can complement each other in two senses. First, they may be used to optimize different infrastructure elements based on their applicability. For example, software with source deduplication capabilities may be used for remote office data protection, storage systems with deduplication capabilities may be used as a backup target for enterprise data center data protection, and compression may be used to reduce the storage requirements for active data. Second, some of the techniques can be used together to achieve additive benefits. For example, compression can be applied to data that has been capacity optimized by other space reduction techniques to gain additional space savings.

Note that some technologies require considerable attention when used together. The sequence in which each technology is applied is important. For example, space reduction techniques may achieve little, if any, benefit when applied to data that has previously been compressed or encrypted. However, compression can reduce space consumption when applied to data that has been single-instanced or deduplicated. Other techniques such as de-constructing compound file formats, re-generating scaled images, and packing many small files into a space allocation unit may also be used to save space.

To achieve complementary benefits, data must be encrypted after space savings techniques are applied. If encrypted data is transmitted, coordination with space reduction techniques must be used to decrypt, deduplicate, and re-encrypt the data to achieve complementary benefits. Common data models can also be used to understand data deduplicated by different technologies. For example, replication bandwidth consumption is minimized if deduplicated data is not inflated when sent.

Space Reduction Ratios and Percentages

A data deduplication ratio over a particular time period is the number of bytes input to a data deduplication process divided by the number of bytes output. Figure 1 depicts the space reduction ratio relevant in most customer situations which reflects all of the complementary capacity optimization technologies actually used.

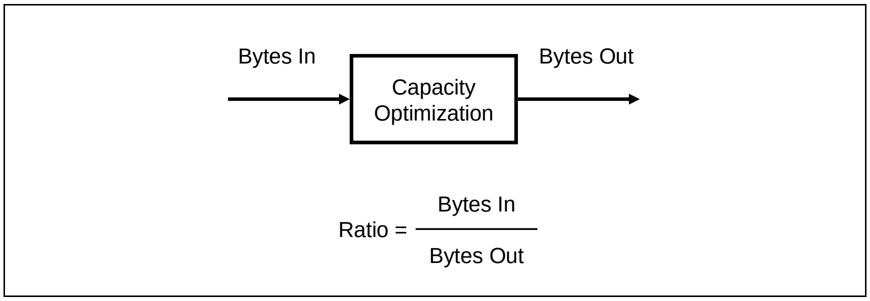

Figure 1. Space Reduction Ratio

Space reduction ratios are typically depicted as “ratio:1” or “ratio X.” For example, 10:1 or 10 X.

The ratio may also be viewed as the data capacity of a system divided by its usable storage capacity. For example, if 100 GB of data consumes 10 GB of storage capacity, the space reduction ratio is 10:1.

Storage capacity is the number of bytes that can be stored on a storage medium. The total or raw storage capacity of a disk drive is based on the number of sectors available for data storage. The usable storage capacity of a storage system is less than the aggregated raw storage capacity, because it accounts for overhead such as spare drives, RAID protection or other form of resiliency, partitioning, file system formatting, and staging areas used by post-processing data deduplication technologies.

Understanding the significance of space reduction ratios requires placing these ratios in perspective:

- Ratios can meaningfully be compared only under the same set of assumptions
- Relatively low space reduction ratios provide significant space savings
- Many factors influence the value of a storage system. Availability, performance, scalability, ease of use, functionality, and affordability factors may outweigh capacity considerations alone.

Comparing ratios is problematic because of the broad set of assumptions implicit in their calculation. The factors that influence space savings will be explored in the next section. Best practice considers these factors and analyzes representative data when evaluating optimizations for specific environments. In practice, many offerings achieve similar capacity savings in specific situations so other factors become differentiating criteria.

It may not be immediately apparent that relatively low space reduction ratios indicate significant space savings. The wide range of space reduction ratios masks the diminishing returns calculated in Table 1 and illustrated in Figure 2. The percentage of space reduction is calculated as 100% less the inverse of the space reduction ratio.

Table 1. Space Reduction Ratios and Percentages

Space Reduction Ratio	**Space Reduction Percentage =** **1 – (1 / Space Reduction Ratio)**
2:1	1/2 = 50%
5:1	4/5 = 80%
10:1	9/10 = 90%
20:1	19/20 = 95%
100:1	99/100 = 99%
500:1	499/500 = 99.8%

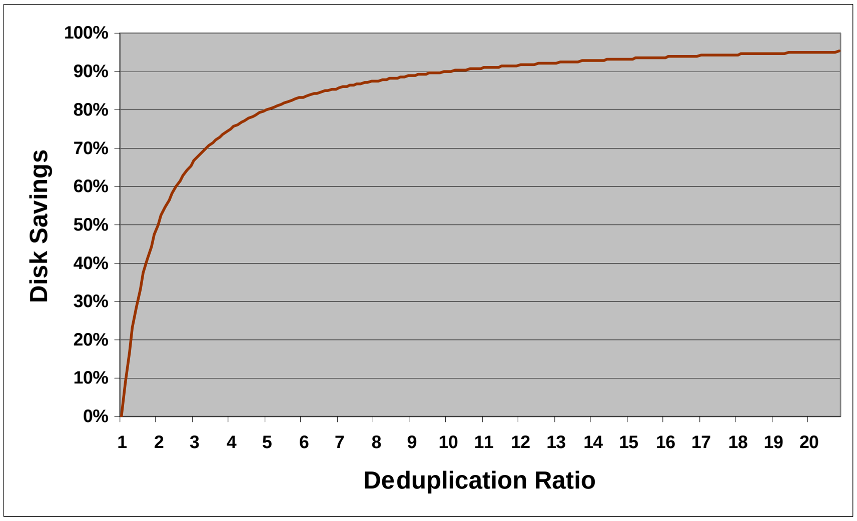

Figure 2. Space Reduction Percentages

The value of a storage system to a business reflects its total contribution to their overall information infrastructure. A storage tier is storage space that has availability, performance, and cost characteristics different enough from other storage tiers so as to economically justify movement of data between it and other storage tiers. Capacity optimization techniques can affect a lower cost per byte for a particular storage tier but it's important to consider data deduplication within an overall ILM context. For example, eliminating unnecessary data is more efficient than simply reducing its storage footprint.

Factors that influence space savings

Data deduplication achieves space savings by finding and reducing the number of duplicate copies of data that are stored. The ability of an implementation to recognize duplicate data determines how much space can be saved but no amount of "secret sauce" can save space if there is no duplicate data. Thus, the amount of duplicate data sets an upper limit on the space savings which can be achieved by data deduplication. The amount of duplicate data in a specific environment is determined by the characteristics and access patterns of the data and by the operational policies and practices in use. The space savings actually achieved depends on which data is deduplicated and on the effectiveness and efficiency of the specific technologies used to perform capacity optimization. Not surprisingly, actual deployment of specific technologies in specific environments determines the actual space savings.

Data characteristics and access patterns

The type of data is a good indicator of how well it will deduplicate. Files created by office workers often contain redundant data and are frequently distributed or copied. At the other extreme, data derived from natural sources is typically unique. Estimating the benefits of data deduplication for a specific environment may be performed using a redundancy modeling or analysis tool which reflects the level of effort and accuracy desired. Note that company policy may influence the type of data that may be stored on business systems; for example, if a company does not backup MP3 or JPEG data or prevents creation of PST files, such data would not be deduplicated.

The frequency that data is changed also impacts the likelihood that duplicate data will be created or detected. The less that data is modified the greater the chance that copies of that data will contain the same data as other copies of that data. Frequent update, copy, or append operations may also make it more difficult for some algorithms to detect duplicate data. For example, the segment boundaries used to deduplicate a virtual machine image may need to change when the file system moves data to different storage allocation units. Generally, the data deduplication ratio will be higher when the change rate is lower. This also implies a higher data deduplication ratio should be expected as the percentage of reference data to total active data increases because reference data is not changed.

As the total amount of data increases there is a good chance that this growth is due to storing data that does not duplicate existing data. As long as the growth is not due to simply making copies of data, a higher growth rate will result in a lower data deduplication ratio because there is more unique data.

Capacity optimization can be applied to both active data and inactive data. Compression may save less space than data deduplication but the results are more consistent across data types and its space savings are additive when performed after data deduplication. Remember that data deduplication usually does not save space when applied to pre-compressed or encrypted data because such data is unique by design.

Data protection applications repeatedly create copies of data so data deduplication techniques have been rapidly adopted to improve the economics of backup to disk. Keeping more data online longer opens up new opportunities to reduce business risk and expand revenues. Data deduplication thereby enables enhancements to recovery, regulatory compliance, legal discovery, and business analytics.

Applications with high data transfer rates require significant processing to deduplicate data as it is being processed by the application, resulting in index size and performance challenges. One approach to meeting this challenge is to use methods which reduce the compute load by reducing the ability to detect redundant data. This approach typically results in a reduction in the data deduplication ratio.

Operational policies

A non-intuitive observation is that widely varying data deduplication ratios can result in identical amounts of deduplicated data, even when the same data deduplication technology is used. The reason is that the numerator (bytes in) can change while the denominator (bytes out) is constant. This arithmetic truism is demonstrated by discussing commonly implemented ***backup methodologies***.

Full backups save a complete copy of a data set each time a backup is performed and results in storing a large amount of duplicate data over time. Differential incremental backups only copy data that was modified since the last full backup or differential incremental backup; this reduces the amount of data sent but requires the latest full backup and *all* newer incremental backups to restore the data. Cumulative incremental backups take a middle ground by copying data that was modified since the last full backup so only the latest full backup and *the latest* cumulative incremental backup are required to restore the data.

The amount of data stored after a backup is deduplicated is identical regardless of which backup methodology is used with a given deduplication technology, thereby holding the denominator constant. The numerator of the data deduplication ratio however will vary widely depending on the backup methodology chosen. For example, full backups of 100 GB over 7 days where the amount of unique data is 20 GB, would result in a data deduplication ratio of 35X (700/20). However a full backup plus six incrementals (each with 3 GB changed data) on the same data would result in a data deduplication ratio of 5.9X (118/20).

While always using full backups with data deduplication offers the fastest restore time (because no incremental backups are required for restore) the amount of data transferred over the network is an important consideration before changing backup methodologies. If backup data is deduplicated before transferring over a network there is strong motivation to always use full backups.

The length of time that data is retained impacts data deduplication ratios in two ways. First, if more data is examined when deduplicating new data, the likelihood of finding duplicate data is increased and the space savings may increase. Secondly, if a data deduplication ratio is calculated over longer periods of time it may increase because the numerator tends to increase more rapidly than the denominator.

Continuing with the previous example, if the full backup data deduplication ratio is calculated over four weeks and 5 GB of additional unique data is stored, the ratio would increase to 112X (2800/25). An even higher data deduplication ratio is calculated if only the last full backup is considered. For example, if the last full backup only stores 200 MB of additional unique data, the data deduplication ratio would be calculated as 500X (100/.2).

The wide range of results calculated in these examples illustrates the importance of understanding exactly what values are used for the numerator and denominator of a data deduplication ratio.

Other considerations have an indirect effect on the data deduplication ratio by influencing the choice of whether or how to deduplicate data. Examples of such considerations include:

- ***Constrained backup windows*** may influence the choice of whether to deduplicate data, deduplicate the data immediately during a backup, or to delay deduplication until after the backup has completed successfully. The decision of how to best satisfy this constraint will result in the selection of a particular set of technologies to optimize storage capacity.

- ***Replicating deduplicated data*** allows more data to be copied to a remote location more frequently using the same network to improve disaster recovery RPO/RTO service levels. If a data deduplication technology is chosen to ensure data is replicated within a particular window, this selection of a particular technology will influence the data deduplication ratio. For example, if data must be replicated immediately after a backup, then data deduplication technologies that perform deduplication immediately, rather than as a post-processing activity, may be a better fit.

- ***Existing investments*** in skills, processes, and equipment may also influence whether deploying data deduplication in specific environments is economically justified. For example, if remote office data protection requirements are being fulfilled, there may be little incentive to disrupt current practices. On the other hand, if remote office data protection is believed to represent a significant exposure, data deduplication can reduce the cost of managing this business risk.

Space reduction technologies

The effectiveness and efficiency of the data deduplication process also depends on the technologies used to perform space reduction. Decisions involving where, when, and how to deduplicate data come into play once the selection of which data to deduplicate has been made.

The location ***where*** data is deduplicated influences the data deduplication ratio insofar as it changes the scope of data examined during the space reduction process. There is a greater chance that duplicate data will be found as more data is examined. ***Source and target data deduplication*** technologies use different approaches to deduplicate data across multiple locations. Both approaches can optimize bandwidth when transmitting deduplicated data between locations.

Figure 3 illustrates multiple source data deduplication engines using a common deduplicated data repository to reduce duplication across multiple locations. This approach is typically implemented using software agents at each source location.

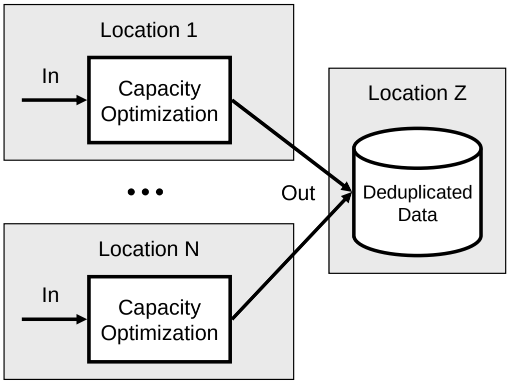

Figure 3. Source data deduplication

Figure 4 illustrates multiple target data deduplication engines replicating deduplicated data from multiple locations to a common deduplicated data repository. The replication and data deduplication processes are integrated to avoid redundancy across locations. This approach is typically implemented using hardware appliances at each location.

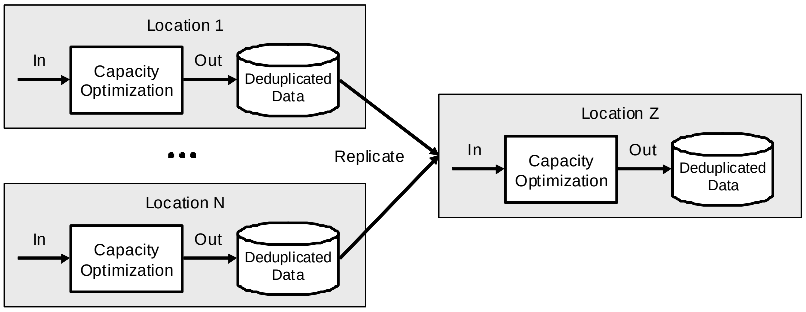

Figure 4. Target data deduplication with replication

In general, the wider the scope of data deduplication, the higher the data deduplication ratio will be. Global deduplication technologies support deduplication of data within multiple storage systems that may span locations. Local deduplication technologies support deduplication of data stored within a single storage system.

The relative time ***when*** data is deduplicated (***immediately or as delayed activity***) only has an indirect influence on the data deduplication ratio. The computational requirements to deduplicate data as it is being accessed by applications with high transfer rates may influence the ability to detect redundant data because a larger segment size (see below) may be used to reduce the processor load. Alternatively, data deduplication can be performed after the application has completed processing.

Methods for ***how*** data deduplication is performed can be categorized as being hash-based or delta-based. The method selected influences the amount of redundant data that can be detected and reduced.

Hash-based data deduplication, the predominant method, defines and identifies segments of data using hash functions, provides a mechanism to find subsets of the data that are duplicated, and reduces or eliminates this redundancy. It is referred to as block-level because the segments of data that it defines and identifies with a hash are also referred to as blocks.

Large segment sizes tend to decrease the data deduplication ratio because there is a greater chance that some of the data in the segment will differ from that contained in other segments. This is also why sub-file data deduplication provides greater space savings than file-level single-instance storage. However, using smaller segments does require managing more metadata.

Methods which can vary the size of the segment to detect redundant data whose position has been shifted in a byte stream, tend to increase the data deduplication ratio. Methods which use fixed-length segments will not recognize redundant data if it is not aligned on the fixed-length segment boundaries.

Delta-based data deduplication stores or transmits data in the form of differences from a baseline copy. The baseline is a complete point-in-time copy of the data used to recreate other versions of the data. Although not required, delta-based data deduplication may be performed at a byte-level.

The effectiveness of hash-based and delta-based data deduplication methods can also be impacted by the ability to recognize duplicate data when unique metadata is inserted into a data set or I/O stream. For example, unique metadata is typically inserted within database, email, and backup stream data sets. Data deduplication ratios may be increased when the deduplication method is aware of this content. Other ***content-aware*** techniques can also be used to improve space savings as mentioned on Page 4.

Spatial data deduplication refers to the ability to detect and reduce redundant data across different files. Some implementations support data deduplication for files that reside within a single file system and other implementations support data deduplication across multiple file systems spanning locations.

Temporal data deduplication refers to the ability to detect and reduce redundant data from the same file at different points in time. The potential for redundant data increases as the amount of data and the number of files increase. Typically, more redundant data can be reduced from multiple points in time for a file than can be achieved between different files so data deduplication ratios are often higher for temporal than for spatial data deduplication. Hash-based and delta-based data deduplication provides temporal data deduplication; hash-based approaches also provide spatial data deduplication.

Summary

Data deduplication lowers business risks, increases revenue opportunities, and reduces storage tier costs, resulting in a perfect storm for companies deploying an adaptive storage infrastructure. Storage resiliency technologies, such as RAID or RAIN, safeguard the deduplicated data to ensure high availability of applications accessing the data. The economics of data deduplication makes it more than compelling; it is mandatory for any business seeking to maximize their customer service levels.

Data deduplication ratios are easy to over-analyze and attribute benefits to, that may or may not exist.

Table 2 summarizes the factors that influence space savings. Companies are advised to place data deduplication within the perspective of Information Lifecycle Management so the benefits of space reduction are balanced with the performance, availability, usability, maintainability, scalability, and cost characteristics necessary to optimize their data storage investment.

Table 2. Data deduplication ratio factors

Factors associated with higher data deduplication ratios	Factors associated with lower data deduplication ratios
Data created by users	Data captured from mother nature
Low change rates	High change rates
Reference data and inactive data	Active data
Applications with lower data transfer rates	Applications with higher data transfer rates
Use of full backups	Use of incremental backups
Longer retention of deduplicated data	Shorter retention of deduplicated data
Wider scope of data deduplication	Narrower scope of data deduplication
Continuous business process improvement	Business as usual operational procedures
Smaller segment size	Larger segment size
Variable-length segment size	Fixed-length segment size
Content-aware	Content-agnostic
Temporal data deduplication	Spatial data deduplication

The Storage Networking Industry Association (SNIA) Data Management Forum (DMF), Data Protection Initiative (DPI) Data Deduplication and Space Reduction (DDSR) Special Interest Group (SIG), encourages the participation of all parties interested in data deduplication technologies. This paper has focused on ***data*** deduplication technologies. Further investigation into higher levels of abstraction may result in higher deduplication ratios over time and enable the pervasive reuse of ***information*** throughout its lifecycle. For further information related to data deduplication terminology, refer to the DDSR Terms and Definitions Glossary.

About the SNIA

The Storage Networking Industry Association (SNIA) is a not-for-profit global organization, made up of some 400 member companies and 7,000 individuals spanning virtually the entire storage industry. SNIA's mission is to lead the storage industry worldwide in developing and promoting standards, technologies, and educational services to empower organizations in the management of information. To this end, the SNIA is uniquely committed to delivering standards, education, and services that will propel open storage networking solutions into the broader market. For additional information, visit the SNIA web site at www.snia.org.

About the Data Protection Initiative

The Data Protection Initiative (DPI) was created by the Data Management Forum (DMF) to allow industry leaders and participants to come together in a community to focus on defining, implementing, qualifying, and teaching improved methods for the protection and retention of data and information. The DPI operates as an online virtual community, sharing work efforts, training programs, and outreach services such as research, whitepapers and training, and educational courses. This group's primary objective is to serve IT professionals by creating certification, education, and training programs, and by creating a world-class collaborative information portal with global influence and reach. The DPI's goal is to build the knowledge base and training capabilities to become the worldwide authority on data protection, helping all of the SNIA's constituents (vendors, IT, regulatory agencies, and channel partners) to better understand and implement data protection solutions. For more information on the DPI visit http://www.snia.org/forums/dmf/programs/data_protect_init/

About the Author

Mike Dutch is Co-Chair of the SNIA DMF DPI Data Deduplication and Space Reduction Special Interest Group and a Senior Technologist within the EMC Storage Software Group CTO Office. Relevant constructive comments may be addressed to ddsr-sig-chair@snia.org.