

Understanding data de-duplication ratios

Data de-duplication has obvious benefits, but vendors' claims regarding data de-dupe ratios have led to confusion.

BY MIKE DUTCH AND
LARRY FREEMAN

Data de-duplication plays a vital role in helping to manage today's prolific growth of data. Optimizing data storage systems with de-duplication can be part of a broader strategy to provide an efficient data storage infrastructure that is responsive to dynamic business requirements.

According to the Storage Networking Industry Association (SNIA), "Data de-duplication is defined as 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 de-duplication from other space

reduction techniques is the reduction of duplicate data at the sub-file level.

This article explores the significance of data de-duplication space savings ratios, and provides an understanding of various claims made by vendors.

Data de-dupe ratios

Simply stated, a data de-duplication ratio refers to the number of bytes input into the de-duplication process divided by the number of bytes output from the process. The **> p. 16**

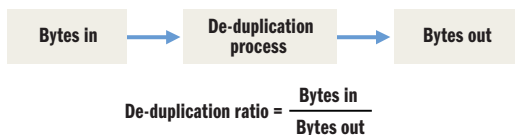
figure below depicts how data de-duplication ratios are calculated.

De-duplication savings are typically depicted as a ratio; for example, 10:1. Understanding the significance of de-duplication ratios also requires the understanding of a few basic points:

- Relatively low de-duplication ratios can yield significant space savings;
- Ratios are meaningful only when compared under the same set of assumptions;
- De-duplication ratios are influenced by many characteristics of the examined data.

Relatively low de-duplication ratios can yield significant space savings. The wide range of space reduction ratios reported by vendors can mask the value of smaller ratios. As shown in the figure below right, relatively low space reduction ratios can still yield significant savings. De-duplication savings ratios of 2:1 or 3:1 may seem unimpressive,

Calculating de-duplication ratios



but when placed in the context of removing 50% or 66% of physical data stored, the benefit becomes much more apparent.

Conversely, once de-duplication ratios become 10:1 or higher, the relative amount of data reduction becomes less of a factor as a percentage of overall data reduction. When claims are made of 50:1 or even 500:1 de-duplication ratios, a dose of common sense is prescribed. For example, de-duplication ratios of 50:1 and 500:1 yield an incremental 8% and 9.8% disk savings, respectively, beyond the 90% disk savings achieved with a 10:1 de-duplication ratio.

Ratios are meaningful only when compared under the same set of assumptions. Data can be generally categorized as either temporal or spatial. Temporal data accumulates over time based on recurring events; an example of this would be nightly backups. Spatial data tends to be more fixed in size, examples being data stored by primary data

applications and within archived volumes.

With de-duplication of temporal data, the potential for redundant data increases as events cause the amount of data to increase over time. This may be particularly true when performing operations such as data backups. Backup data sets often contain a large amount of redundant data. In this environment, de-duplication ratios can be very high.

Spatial data de-duplication, while typically offering lower space savings ratios, still provides a valuable function by reducing redundancies across a broad spectrum of data sets and storage tiers. De-duplication ratios in these environments may be lower, but as previously mentioned still provide valuable space savings. The figure on page 33 illustrates the effect of de-duplication with temporal and spatial data.

De-duplication ratios are influenced by the characteristics of the data. The length of time that data is retained usually

impacts data de-duplication ratios. As more data is accumulated, the likelihood of finding duplicate data is increased and the space savings will thus increase. This is true for both temporal and spatial data:

- The *wider the scope* of data de-duplication, the higher the data de-duplication ratio will likely be. For instance, global de-duplication technologies are available that support de-duplication of data across multiple storage systems that may span several locations.
- The *type of data* can be a good indicator of how well it can be de-duplicated. For example, files created by humans often contain redundant data and are frequently distributed or copied. Examples are application data, such as documents, spreadsheets, and presentations. On the other hand, data created from non-human sources is typically random and unique. Examples include images, audio files, and scientific data acquisition. To help

predict the effect of de-duplication savings on a particular data set, redundancy modeling and analysis tools are available, which reflect the level of data redundancy in a given environment.

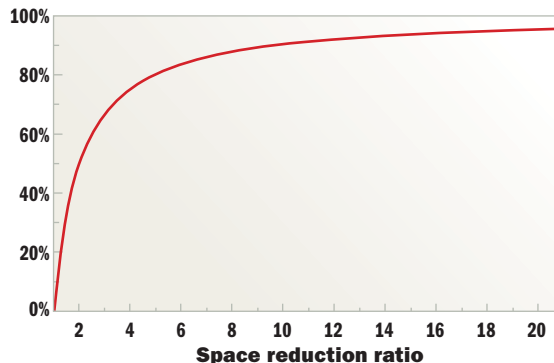
- The *frequency that data is changed* also impacts the likelihood that duplicate data will be created. The less that data is manipulated, the greater the chance that copies of that data will contain the same data as

The economics of data de-duplication makes it more than compelling; it is becoming mandatory.

other copies. Frequent update, copy, or append operations may also make it more difficult for some algorithms to detect duplicate data. Generally, the data de-duplication ratio will be higher when the file change rate is lower. This also implies that a higher data de-duplication ratio should be expected as the percentage of reference data to total active data increases since reference data is not changed.

- The *total amount of data growth* does not necessarily mean that de-duplication ratios will also rise. Growth may be due to storing new data that does not necessarily duplicate existing data. A high data growth rate may actually result in a lower data de-duplication ratio if this growth results in more unique data. ▶ p. 33

Space reduction ratios and percent savings



time. This brought performance with four VMs to just over 160,000 IOPS. Once again, for the read component of the data that we were moving, we had exceeded the capabilities of a 4Gbps HBA.

Following our multiple VM scalability tests, we ran tests on the scalability of a multiprocessor VM with multiple drives. With each disk given its own virtual SCSI adapter, I/O performance with a single VM with multiple drives scaled identically to the I/O scaling exhibited with multiple VMs. Nonetheless, with respect to IOPS, four drives on one virtual SCSI adapter did not scale as well.

Next, we utilized random 8KB I/O blocks, which typify the I/O transactions found in database-driven applications. With 8KB blocks, we continued to put significant stress on both the VSC-VSP mechanism and the QLogic QLE2560 HBA to make transactions; however, we also doubled the data throughput.

By doubling the amount of data per request,

we crossed a threshold not crossed with 4KB requests—we saturated the read channel of the 8Gbps HBA. With just three drives in our initial Windows Server 2008 scalability tests, we exceeded 122,000 IOPS and reached a total throughput of 960MBps. As a result, adding a fourth disk produced only a 5% improvement in IOPS and total data throughput in MB per second. More importantly, all I/O—both reads and writes—coming from the QLogic HBA and measured at the Texas Memory array was perfectly balanced across all four logical disks.

We measured that same pattern in all of our VM scalability tests. With 8KB requests, the ultimate rate-limiting factor was the 8Gbps HBA. As a result, we began to see our scalability tests converge to just under 129,000 IOPS and just over 1,000MBps of data throughput.

In our final tests, we used 64KB I/O blocks, which are found in business intelligence applications, such as on-line analytical processing

(OLAP), data mining, and data warehousing. I/O in these applications is the antithesis of I/O in messaging applications—there are a limited number of users, and the speed at which large volumes of data can be moved dominates in importance. Now data throughput totally dominated our tests, which were identical in all cases as two drives or two VMs were enough to saturate reads on our 8Gbps fabric at wire speed.

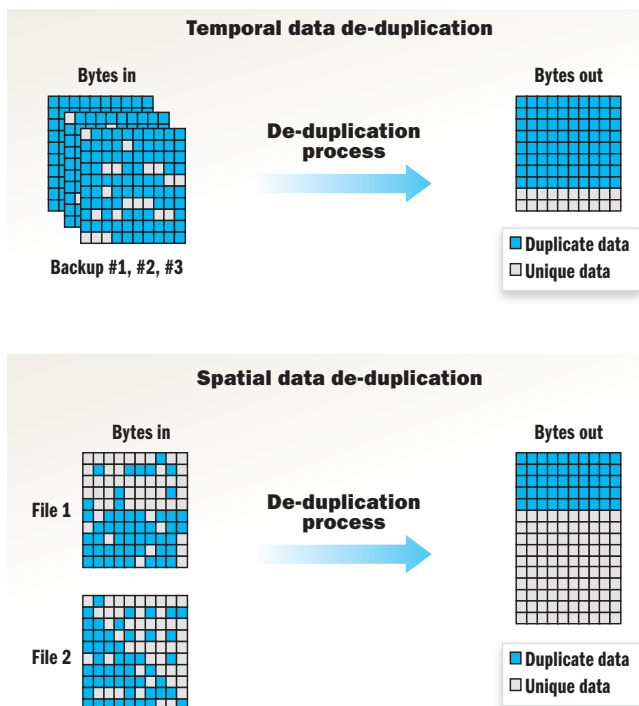
The number of IOPS sustained in all of our tests clearly indicates that a Hyper-V VOE, based on an 8Gbps SAN infrastructure, is able to scale and support a high number of VMs, which will easily provide for a high consolidation ratio. Equally important, the scalability of this SAN infrastructure enables VMs to host the most I/O-intensive applications. ©

VENDORS MENTIONED

Dell, Intel, Microsoft, QLogic, Texas Memory Systems.

Understanding data de-duplication ratios continued from page 16

Temporal vs. spatial data de-duplication



A must-have

A thoughtful deployment of data de-duplication reduces storage costs and allows data to be retained for longer periods of time on disk without excessive physical storage penalties. The economics of data de-duplication makes it more than compelling; it is becoming mandatory for any organization seeking to maximize its data management support levels.

Data de-duplication space savings benefits are sometimes difficult to evaluate due to the widely varying data characteristics from one user environment to the next. A good rule of thumb when evaluating the de-duplication process is to compare the ratio of *bytes input* to *bytes output* in order to calculate the true space savings effect.

Once an accurate calculation is made, an informed decision can be made concerning trade-offs in performance and cost when measured against the potential space savings provided through data de-duplication. ©

For more information on this topic, refer to the SNIA white paper, "Understanding Data De-duplication Ratios" (www.snia.org/forums/dmf/knowledge/white_papers_and_reports/Understanding_Data_De-duplication_Ratios-20080718.pdf).

MIKE DUTCH (EMC) and LARRY FREEMAN (NetApp) are co-chairmen of the SNIA Data Management Forum's Data De-duplication and Space Reduction Special Interest Group (DDSR SIG). The group's mission is to bring together a core group of companies that will work together to publicize the benefits of data de-duplication and space savings technologies. For more information: www.snia.org/forums/dmf/programs/data_protect_init/ddrsig