NetApp™
Go further, faster

This article was originally published in NetApp's Tech OnTap Newsletter.
To receive the newsletter monthly and enjoy other great benefits, sign up today at www.netapp.com/us/communities/tech-ontap

How Safe Is Deduplication?

By Larry Freeman

LARRY FREEMAN
SR. MARKETING MANAGER FOR STORAGE EFFICIENCY
NETAPP

During his 25-plus year career in data storage, Larry has held various positions with companies including Data General, Telex Computer Products, NEC Information Systems, and Spectra Logic.

A frequent speaker and author, Larry's current role at NetApp is evangelizing data storage efficiency technologies, including deduplication. Larry is the founder and co-chair of the SNIA Data Deduplication and Space Reduction Special Interest Group and active in the SNIA Green Storage Initiative.

RELATED INFORMATION

- Choosing the Right Backup Technology (www.netapp.com/us/communities/tech-ontap/backuprecovery-0608.html)
- Top Three VMware Backup Challenges (www.netapp.com/us/communities/tech-ontap/backupvm-0608.html)

Unless you've had your head in the sand recently, you're probably well aware that deduplication is hot. It seems like every storage vendor you've ever heard of (and many more you haven't) is touting deduplication technology as a way to reduce the cost of disk-to-disk backup.

The question you're probably asking yourself is whether deduplication of your data is safe. When the time comes to restore data from that deduplicated backup, will you actually be able to get your data back?

Assessing the relative safety of a deduplication technology really comes down to two basic components:

- The algorithms that are used to identify and eliminate duplication
- The reliability of the underlying hardware and software

In this article, I'm going to take a look at deduplication technology in light of these two criteria. I'll also explain the choices that NetApp has made to enhance the reliability of our own deduplication technology. Because we support deduplication for both primary and secondary storage—whereas most other vendors provide deduplication only for backups—data safety is of the utmost importance for us.

IDENTIFYING DUPLICATE DATA

Most existing deduplication products operate at the block level—new blocks are compared against previously stored blocks to determine whether an identical block has been previously stored. If it has been previously stored, the "new" block is discarded in favor of a pointer to the stored block.

So how do you determine whether two blocks are identical? The most common method is that for each block, you compute a "fingerprint," which is a hash of the data contained in the block. If two blocks have the same fingerprint, they are usually assumed to be identical.

However, there is a small but nonzero chance that two nonidentical blocks will yield the same fingerprint or hash value. This is termed a "hash collision" and can result in a unique data block being accidentally deleted.

As you might expect, reducing the probability of a hash collision requires a more complicated algorithm, which typically consumes more CPU resources to compute the hash and yields a larger output value. So there's an obvious trade-off between reliability and speed. Also, longer hashes consume more space for fingerprint storage.

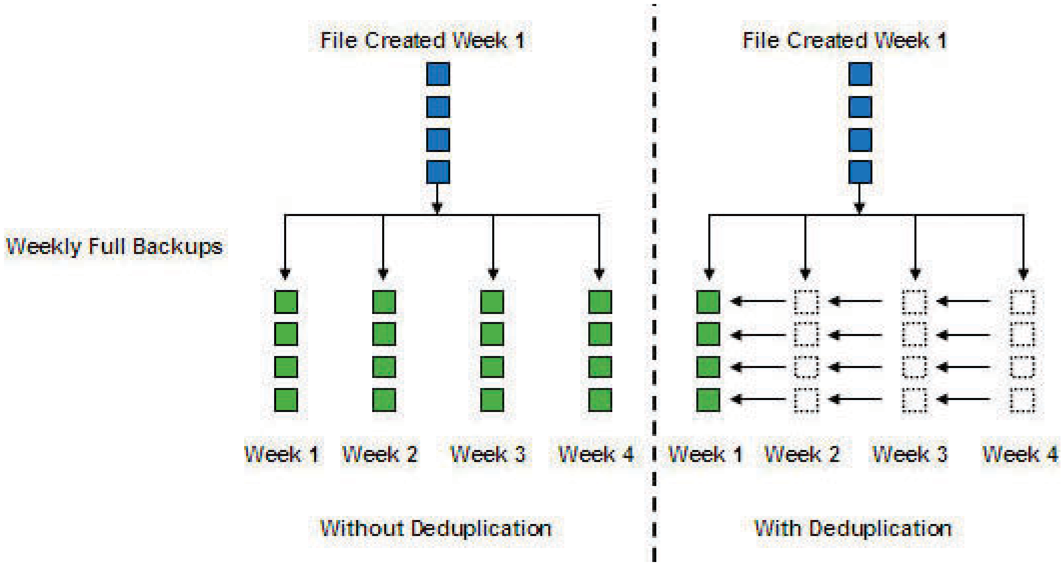

Figure 1) Four weeks of full backups of the same unchanged file. Without deduplication you have four discrete copies of the same file. With deduplication you have a single copy. This yields substantial space savings, but at the same time it makes reliable storage of that single copy more critical.

TOUGH QUESTIONS, REAL ANSWERS

What kind of real-world questions are people asking about deduplication? View a "no holds barred" transcript from a recent live chat (http://www-download.netapp.com/edm/TT/docs/20080522COD_transcript.pdf)

MANAGE DATA GROWTH WITH DEDUPLICATION

A recent TechTalk Webcast zeroed in on some of the biggest problems in backup and recovery today, including:

- How to cut costs with deduplication
- Low bandwidth WAN backup
- Making the move from tape to disk-based backup

Watch the Webcast to learn more about these important topics (http://www-download.netapp.com/edm/TT/WOD/WOD20080605/index.html).

When you are evaluating deduplication technologies, you need to find out how a vendor identifies duplicates and ask about the risk of hash collisions with the chosen algorithm. Many vendors will argue that the chance of a hash collision is lower than the probability of a disk failure or a disk drive error or tape error that corrupts a data block. I don't know if that's a truly comforting thought or not, but I believe that most of us would prefer to minimize as many risks as possible.

NetApp supports deduplication for both primary and backup storage, we take a more aggressive approach to preventing hash collisions. We use a fingerprint algorithm like most everyone else, but we use it only to identify potential duplicates. When that happens, we do a byte-by-byte comparison of the two blocks to check that they are identical before discarding any blocks. NetApp technical director Blake Lewis provided a more detailed explanation of how NetApp identifies duplicate blocks in a previous article (http://partners.netapp.com/go/techontap/dedupe.html).

RELIABILITY OF UNDERLYING HARDWARE AND SOFTWARE

Deduplication is only as reliable as the underlying hardware and software. In fact, although it may not be immediately obvious, with deduplication, reliability becomes even more crucial.

For example, suppose that you run a fairly standard backup schedule with nightly incrementals and weekly full backups. Now suppose that you create a file at the beginning of the month and then make no changes to it. With traditional backups, you'll have four copies of the file at the end of the month, one for each week's full backup. If you need to restore the file at that point, the probability is high that you'll be able to restore at least one of the four copies, even if your backup medium is not reliable.

But when you bring deduplication into the picture, at the end of the month you have only one physical copy of the file—the copy created by the first full backup—plus three sets of pointers to the same file blocks. Based on this simple example, it should be clear that you want to be sure your deduplicated backups have been reliably stored on resilient hardware with good RAID protection. Over the course of a year you could have hundreds of backups that actually reference most of the same data blocks.

There are a wide variety of deduplication products out there. Some are software only and may use a variety of underlying hardware; some include both hardware and software (possibly obtained from a variety of sources through licensing or OEM arrangements). Before making a decision, you should assess how mature the software is, how robust the underlying hardware is, and how well the two integrate together.

NETAPP RELIABILITY

With NetApp® storage, deduplication is an integral part of the Data ONTAP® operating environment that runs across our entire product line. Data ONTAP has been under continuous co-development with NetApp hardware platforms for more than 15 years. Unique features of the NetApp WAFL® technology actually simplify the implementation of deduplication and make it possible to deduplicate any stored data, not just backup data.

Proven reliability features in NetApp hardware and software result in data availability of more than 99.999% as measured across the NetApp installed base. A recent analyst report describes the NetApp methodology and many of the features that contribute to NetApp's reliability (http://media.netapp.com/documents/ar1056.pdf).

One example of our attention to detail involves the well-known fact that disk drive bit errors can develop over time—or even during the manufacture of disk drives. Every drive has built-in error correction that detects and usually corrects such bit errors. If a string of errors is too great to be handled by ECC, the drive reports back that the sector is unreadable, at which point RAID algorithms fix the error from the information stored on other sectors. NetApp, however, also uses a checksum scheme for further protection—we use an additional portion of the drive as overhead to store checksums that move with the data through the system to check that what was written is returned perfectly during data restoration. In essence, we provide a third level of protection.

To protect the reliability of data committed to disk, NetApp also developed RAID-DP™, a high-performance, dual-parity RAID 6 implementation that protects against double disk failures without sacrificing write performance. You can read more about RAID-DP and other NetApp enhancements to protect against misbehaving disk drives in a previous Tech OnTap article (http://partners.netapp.com/go/techontap/matl/sample/0206tot_resiliency.html).

CONCLUSION

To protect your backup data, deduplication technology must use appropriate algorithms to avoid discarding unique data blocks and also provide the fundamental hardware and software reliability necessary to safely store deduplicated data for later recovery.

Because NetApp deduplication technology is used for primary data stores as well as for backup data, we take extra care to protect data reliability. NetApp deduplication uses a combination of fingerprints plus byte-by-byte block comparisons so that unique data blocks are never erroneously deleted due to hash collisions. Deduplicated data is stored on NetApp storage systems using hardware and operating software that have been proven reliable and resilient through years of field deployment, so you can be confident that when it comes time to recover data, you'll get back the data you backed up.

Tech OnTap

SUBSCRIBE NOW

Tech OnTap provides monthly IT insights plus exclusive access to real-world best practices, tips and tools, behind the scenes engineering interviews, demos, peer reviews, and much, much more.

Visit www.netapp.com/us/communities/tech-ontap to subscribe today.

QUESTIONS ABOUT DEDUPE?

Want to learn more about NetApp deduplication? Larry Freeman hosts a forum on the topic on the NetApp communities site. You can read past postings, share your experiences, ask questions, and get feedback from other users (http://communities.netapp.com/community/our_products_and_solutions/deduplication).

ESG REPORT ON NETAPP DEDUPLICATION

In a recent independent study, ESG evaluated the effectiveness of NetApp deduplication in terms of both capacity savings and performance impact.

Environments evaluated included:

- Virtual server storage
- Backup
- Archiving

Read the full report to find out the results (http://www-download.netapp.com/edm/TT/docs/ESG_Lab_Validation_Report.pdf)

NetApp creates innovative storage and data management solutions that accelerate business breakthroughs and deliver outstanding cost efficiency. Discover our passion for helping companies around the world go further, faster at www.netapp.com.

NetApp™
www.netapp.com