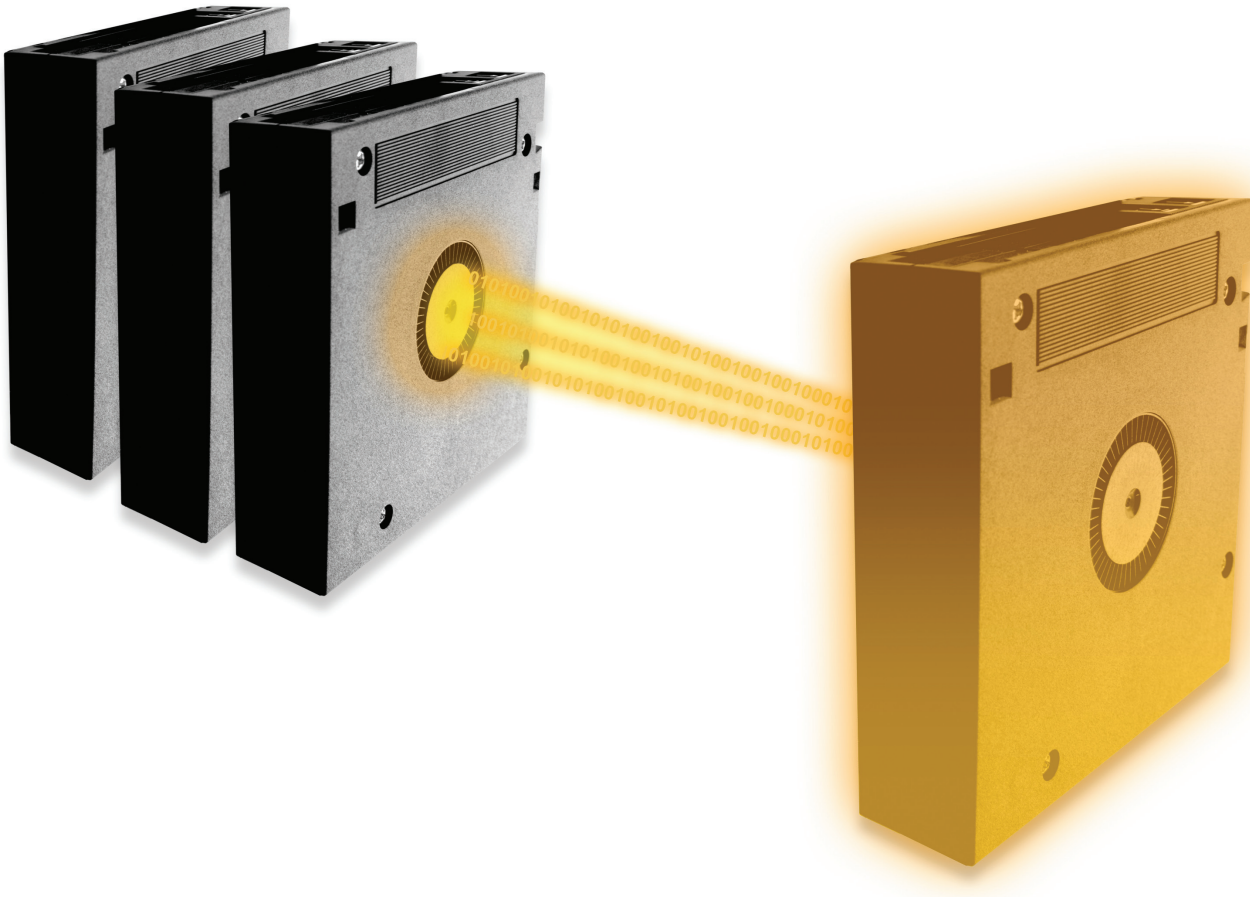


# TAKING A LOOK AT DEDUPLICATION

BY ROB SIMS



We've all seen data deduplication quickly evolve as one of the hottest topics within the industry. And why wouldn't it be? After all, any technology that eliminates unneeded copies of the same identical data would seem to be a winner – in fact, a major winner from nearly every perspective. At first glance, data deduplication may appear to serve as an extremely handy trash compactor that greatly reduces costs and the amount of data an organization stores with minimal impact to your standard business activities.

While there are a few fitting scenarios where deduplication is perfectly suitable for an organization, it is not a wonder drug that fixes all of your current challenges and requirements. Let's take a deeper look and carve out specific issues your company should address while considering deduplication, which for this article is considered a data storage approach eliminating redundant blocks of data such that duplicate blocks are not stored but rather referenced by a list of pointers using an index.

If you decide to deploy this technology – especially on the systems side, you will want to consider the following seven problematic issues:

## 1 Deploying a technology still in its infancy: Are you willing to put the company on the line for data deduplication?

Whenever a new technology such as deduplication doesn't meet specific standards, you are going to quickly see serious complications arise. Deploying a new technology during its infancy can often create significant risk. Every vendor has a different method, and therefore, there is no interchange or interoperability for the end user. What if a vendor goes out of business or is purchased? In that case, there is no viable path for taking the data from one vendor and moving it to another. The user will be locked into that vendor.

For example, a company that uses deduplication from vendor A acquires another company of similar size, but that company uses deduplication from vendor B – they are now faced with either maintaining both vendor relationships and keeping the data separated, or storing all new data on one of the systems. This of course will effectively double their cost and take away access to the other perfectly good deduplication system.

## 2 Data restore the real issue at hand: How long can your organization survive without business critical data?

Placing aside the relative youth of deduplication, the objective for which it was created may often conflict with the most basic of all data storage goals: data restore. So much focus on data backup is placed within most discussions of deduplication that organizations often forget that the only reason we backup data is to have the ability to restore it.

What many don't prioritize as a concern is the extreme costs of data restore. As a deduplication volume continues to grow, the time for restore also grows. Data restoration from tape once it is loaded is by far the fastest restore method, but there are many organizations starting to see the penalty they are taking for their "trash compacting" when it comes time to pull the data back out and save their data (and their business) at a critical moment. This isn't the case for small restore needs, but the pain is felt considerably during full system restores.

A deduplication system has to maintain the hash tables/database and process every data block that comes to the appliance. Every block requires the same amount of processing. Therefore, the system is running at full tilt for every data transfer (much worse for data restore as the search for all of the correct blocks must be done first before putting the data back together to transfer to the host). Usually, performance of restore is critical to organizations. It's still too early to tell if this is becoming a major hurdle for deduplication.

## 3 Interoperability with physical tape storage: Will your IT department always operate in a technological silo?

Many organizations are realizing that backing up data to tape still stands as a cost-effective and secure method for long-term storage. Therefore, they would like to backup the data from the deduplication system to physical tape much like a virtual tape library does. This frees up the backup process such that the virtual tape library can send the data to tape more efficiently (these are the tenants of hierarchical storage). Since deduplication engines must reconstitute the data to write to tape, they run into the problem discussed earlier with data restore. This is a slow process such that organizations don't gain any performance benefits writing to tape. In fact, it is worse than if the data were written to tape directly. This is what many are starting to realize such that they are using deduplication for short term backup, but writing in parallel to tape for long-term and disaster recovery protection.

## 4 The safety of long-term disk-based storage associated with deduplication: Does your company have requirements to protect data stored on a long-term basis?

Deduplication brings to the table a long-standing misconception that disk-based data storage is safer than tape. And why is this a relevant issue? As the data being stored in the deduplication system grows, and as the stored data ages, one very important point an organization must address is data safety. When the deduplication engine receives data, it performs the hash algorithm and compares to the existing hash. Any new data is then stored and the hash table/database is updated. All subsequent data transfers are compared to the hash – not the actual data. The actual data is sitting idle until it is read. Therefore, as the system ages, there could be entire blocks of data that have never been read. This speaks to the inherent errors with magnetic disk.

Organizations are often told that disk is more reliable than tape. But that is simply not true. Disk systems use RAID to deal with the inherent failure rates. SATA disks are even worse. Most deduplication systems use SATA disk arrays for data storage. Unless the actual data is refreshed and/or restored, there is a growing "time-bomb" whereby the data stored might not be readable as the SATA system ages and fails. Instead of using just one disc as a redundant disk, use RAID 6 (which uses two disks) – and requires more management, disks and processing power. In order to avoid major data losses, RAID 6 becomes an absolute requirement. As organizations are looking to increase their redundant processing capabilities, we are likely to see IT departments caught in this spiral growth requiring bigger disks, more time, and much more energy.

CONTINUED ON PAGE 72

CONTINUED FROM PAGE 71

## 5 Data deduplication is tied to large energy consumption and bills: Is your IT department asked to practice green IT ?

Essentially every new data set must go through the complete hashing algorithm and data comparison process within a deduplication system. There is no economy of scale capable because until the hashing process (the most process intensive component) is complete, the deduplication system cannot tell if the new data happens to be the same data that was just sent or is associated with a new process. The challenge then for the user is that deduplication isn't providing any value add from a data perspective. It is simply providing a bigger trash compactor to squeeze more on top. However, if the data were managed such that duplicate data were only stored and protected once, then the volume of data coming to the deduplication engine would be less, the bandwidth would not be impacted, and the amount of new data would be less.

## 6 Absence of separation of data and duty: Compliance: What are lawyers bound to question?

With deduplication, there is absolutely no separation of data. In fact, the data boundaries are completely erased as the deduplication algorithm carves up the data blocks with no care for the container they're in. This brings issues with Sarbanes-Oxley, which requires financial data to be separated. It also creates issues for outsourced IT providers that might have multiple customers for their "shared" assets.

Undoubtedly, lawyers are going to be lining up to raise integrity and security questions when a company must prove the data produced is actually the data stored. It cannot be done with a deduplication system. Another big security issue surrounds "how can you prove the data hasn't been modified?" By definition, the data inside a deduplication system has clearly been modified.

## 7 Data deduplication is reduced to complete inefficiency through encryption: How does deduplication work with current security practices?

One must encrypt the data before it is deduplicated to follow any standard, including FIPS. Once data is encrypted it is completely random. Therefore, de-duplication lacks much value. Let's use a set of twin brothers as an example. In the case of encryption they would both look completely different in all aspects. Therefore, finding large enough sections that can be used to describe both of them is greatly diminished. This happens to data post encryption with deduplication. Encryption with deduplication has a heavy penalty.

Deduplication is used as a single shoe that is supposed to fit all: but it just does not work for all. On the positive side, data deduplication can reduce costs and resource requirements in simpler data storage environments. As soon as you are contemplating more complex and long-term data storage requirements, you have to seriously reevaluate if you are not going beyond the limits of data deduplication's design and intent.

Organizations should thoroughly investigate the considerations discussed above as there are many over-extended data storage deployment cases where deduplication brings more challenges and costs than benefits. It's a hot topic, and deduplication can be highly-beneficial to your organization as well as be your worst enemy – by rising hidden costs, low efficiency, and resource requirements beyond your corporate capabilities – and could bring down an IT manager's job or a company. Can you afford to take such risks? Trends come and go, but business and legal needs for data availability will not disappear. Measure what your needs are and perform due diligence to ensure solid long-term data storage practices.

### ABOUT THE AUTHOR

Rob Sims is President and CEO of Austin, Texas-based Crossroads Systems ([www.crossroads.com](http://www.crossroads.com)), a provider of solutions to connect, protect, secure, and restore data. He can be reached at [rsims@crossroads.com](mailto:rsims@crossroads.com).

## IT/IS NUGGETS

### FROM GANTTHEAD

#### Resources on Cloud Computing and Virtualization

Many organizations are finding that the flexibility and cost savings of cloud computing and virtualization to be beneficial to their bottom line. A login required to read full article.

#### Basics of Cloud Computing

Cloud Computing is the newest buzzword in the IT world, but is it really all that new?

Here's a quick primer to take away any remaining mystery surrounding this latest technology and to help you decide if it's worth embracing.

#### Looking at IT Governance Through the Clouds

Cloud computing offers cost savings, flexibility and speed of deployment that can be very tempting to all kinds of IT organizations, but don't forget about the risks inherent in turning over control of your data in the age of IT Governance and SOX compliance.

#### Making a Good Case for Virtualization Technologies

Virtualization as a concept has been around since the 1960s. There are still kinks to iron out, but it has proven itself as a sustainable technology that only improves over time. Today, cost-constrained companies are making a good case for virtualization technologies.

Visit [www.GUIDErequest.com/IAS](http://www.GUIDErequest.com/IAS) to link to this articles.