Data deduplication has become a standard offering in backup product portfolios. As a technology that has been mainstream for over five years, some IT decision makers may tend to treat deduplication as an afterthought. In actuality, how data deduplication is implemented is a critically important factor towards determining its ultimate success with solving backup ills in the data center or creating them. This is especially true when considering how deduplication can potentially impact recovery time objectives (RTO).

Dedupe Induced Latency

While deduplication undoubtedly offers benefits from a backup data retention and replication standpoint, it can introduce latency into the backup process itself. As a computational and memory intensive process, deduplication algorithms take time to process before data lands on disk.

Effectively it comes down to a time and resource crunch. As we discussed in our previous article, “How Backup Disk Architecture Impacts the Backup Window”, in an effort to overcome this latency problem, some backup appliances have thrown hardware at the problem in the form of CPU power. But this only increases the cost of disk based backup and makes it less competitive with traditional tape based backup systems that are already in place.

The larger issue and one we will focus on in this article, is the time it takes to recover data from a deduplicated backup copy. The fact is, reconstituting or "rehydrating" deduplicated backup data back to its native format, introduces latency into the recovery process; especially if that recovery is large or has many files associated with it.

If a recovery operation is taking place from a backup image, presumably the information is not available elsewhere. Which means that the affected business application will remain down until the data can be fully recovered. As a result, the data deduplication dehydration/rehydration process can potentially have a detrimental impact on Recovery Time Objectives (RTO).

Hybrid Powered Recoveries

One way to circumnavigate the latency issues that data deduplication introduces into the recovery process, is to implement a hybrid disk backup solution. In other words, a backup platform which incorporates both non-deduplicated disk storage to enable rapid recoveries, and a separate deduplicated storage partition which efficiently stores backup images for longer-term data retention and offsite replication. Think of it as a recovery zone and a backup archive zone.
The non-deduplicated disk storage partition would only need to accommodate a week’s worth of backup images (one full backup and 4 nightly incrementals) since the vast majority of restore requests typically take place within the first 7 days of data creation. While this does require sacrificing some disk storage efficiency, the benefit is that there would be no deduplication induced performance penalty during the restore operation. Furthermore, since only one week’s worth of data would need to be maintained in the native area, this portion of the disk footprint could be reasonably small. The remaining portion of the disk backup system could then be reserved for efficient deduplicated backup images.

Making this small architectural change can have a big impact on how quickly organizations can respond to critical data restore requests; especially large ones like a full virtual machine or server recovery. In fact, the potential time difference between performing a full application restore from native disk and deduplicated storage can be quite dramatic. In some cases, it can be merely a matter of minutes to restore a large data set from non-deduplicated storage versus the number of hours it could take to restore the same information from deduplicated disk. Clearly this can be the difference between successfully meeting RTOs or needlessly extending application downtime on critical business systems.

Backup Image Booting

There are additional practical considerations for implementing a disk backup solution which comprises a separate non-deduplicated disk partition. Virtualization backup applications like Veeam and Dell’s VRanger now provide support for booting virtual machines (VMs) directly off the backup images on disk. The concept, often called instant recovery, allows administrators to point their users directly at VM backup images on disk, to achieve rapid application recovery. Needless to say this is an extremely popular feature and many times a new backup application is selected solely for this capability.

This may be a valuable option for those environments that don’t have high availability clustering solutions deployed, however, if VM operating system backup images and application data are only stored in a deduplicated format, instant recovery could be a non-starter. The time it would take to “rehydrate” deduplicated data before it could be presented back to the VM on the backup platform in its native format could simply take too long.

In short, performing “in-place” data recoveries on a deduplicated disk partition might be untenable due to the latency that is introduced into the process. Blocks of data would need to be continuously un-deduplicated and re-deduplicated. The random I/O would kill performance and the deduplication engine itself. In fact, this performance can be so poor that some deduplication appliance vendors recommend installing stand-alone disk in front of their systems just for the purposes of instant recovery. This would obviously introduce additional cost and management complexity into the backup environment.

On the other hand, if backup images are immediately available in their native format, virtual administrators can conduct both fast recoveries of their virtual systems to primary storage or employ the use of an in-place data recovery technology directly from the backup image. This enhances the value of the investment in the disk backup storage platform as it can now be used as a way to further improve RTOs for critical virtualized application infrastructure.

Conclusion

Organizations that are interested in implementing data deduplication to improve efficiencies may want to consider how deduplication can adversely impact their recovery windows. If improving application RTOs are a concern, then deploying a disk staging landing area without deduplication could prove beneficial.
As importantly, however, the system should also include a separate disk zone for storing deduplicated backup images so that organizations can attain the benefits of efficiently maintaining data on disk for extended retention. It also fosters the efficient replication of backup data to a secondary location for DR purposes.

To simplify operational management and control, technology decision makers may want to look towards unified backup solutions which provide a hybrid approach that constitute the best that native disk and deduplicated storage have to offer.

Disk backup solutions like those provided by ExaGrid, give mid-sized to small enterprise data centers the best of both worlds - fast recovery on a classic disk staging area with a separate data duplication storage zone. As we covered in our article entitled, “How Disk Architecture Impacts The Backup Window”, a separate disk-landing zone also helps reduce backup windows and speeds up data transfer time to tape media. This hybrid approach to backup provides organizations with a way to reduce backup windows, speed up data recoveries from disk and garner all the disk retention and WAN replication bandwidth efficiencies available through data deduplication.

About Storage Switzerland
Storage Switzerland is an analyst firm focused on the virtualization and storage marketplaces. For more information please visit our web site: http://storageswiss.com/

Copyright © 2014 Storage Switzerland, Inc. - All rights reserved