Storage Switzerland, LLC

Overcoming Challenges When Integrating Disk Backup Appliances With Veeam

Prepared by: George Crump, Lead Analyst
Prepared: September 2014

There is a new breed of backup applications that is gaining acceptance in data centers of all sizes, including the enterprise: VM Specific Backup Applications. These solutions were designed from the ground up, with virtualization in mind and they do their best to fully exploit the virtual environment. According to Storage Switzerland's research, this new category of products, led by Veeam, represent the fastest growing segment of the data protection market. The challenge facing IT designers is making sure the rest of the environment can take full advantage of the capabilities of these new software solutions.

In prior articles and webinars, we have written about the most glaring VM specific challenges that products like Veeam place on traditional disk backup appliances, but there is also a whole host of challenges that these solutions introduce when backing up to these appliances. In this article, we will cover the backup challenges and how to overcome them.

The Straight Disk Challenge

The biggest challenge that solutions like Veeam present is that they were designed under the general assumption that their backup target would be standard, inexpensive disk arrays directly (via SCSI) attached to the backup server or via simple iSCSI devices. The problem is that using standard disk arrays can become an expensive proposition when they are used for anything more than storing the most recent backup, due to keeping weeks of retention.

As a result of the need to maintain weeks of retention, most data centers want to use more advanced disk backup appliances that leverage capabilities like deduplication, compression and have the ability to replicate data off-site. These systems simplify and reduce the cost of storing data on disk by only storing the unique blocks or bytes between backups. They also can automate the movement of data between sites providing the foundational component of a disaster recovery plan.

The Full Backup Challenge

The first challenge facing companies that want to use both disk backup appliances and solutions like Veeam is managing full backups. Full backups seem to be ideal for disk backup appliances - they are large, sequential and full of redundant data. There are two problems, however. First, many VM specific backup solutions provide some form of rudimentary source side deduplication to reduce the amount of data transferred, but again, those backup solutions expect a standard disk array as a target. The combination of source side deduplication at the backup software layer and target side deduplication at the disk appliance can increase backup time.

The second problem is that most disk backup appliances connect into the backup architecture via Ethernet and make themselves available by a simple CIFS or NFS connection. The problem is that these protocols are not really optimized for the large sequential transfers that typically make up a full backup. Even if source side deduplication is available, it is often rudimentary and these transfers can still be quite large.

Evidence of this problem can be found simply by investigating the number of backup software solutions that have provided an alternative protocol designed specifically for backups. These protocols typically deliver a significant performance increase over CIFS and/or NFS. Most VM specific disk backup solutions, however, are not mature enough yet to be able to deliver their own proprietary protocol that disk backup appliance vendors can readily adopt.

Changed Block Tracking Backup Challenge

The second challenge facing IT planners that want to combine the use of disk backup appliances and VM specific backup software is dealing with changed block tracking backups. These backups are the antithesis of full backups, in that they are small and can be done a couple of times throughout the day or at least several times in between fulls. In fact, if it weren't for the issues when surrounding synthetic fulls that we discuss in the next section, data protection administrators would love to perform CBT backups throughout the day.

CBT backups fundamentally change the shape of the backup process. Because they only back up blocks of information that have changed since the last backup, not the last full, these backups tend to be small and fast. When performed across a few dozen virtual machines, the I/O pattern traveling across the network more closely resembles database I/O than it does backup I/O, in that it is small and seemingly random. Even though changed block tracking can reduce the backup window the result on disk is about a 2:1 deduplication rate. As retention grows, the amount of disk required grows rapidly.

To deal with this I/O many customers will establish multiple Veeam servers to act as data movers, when these servers all start sending data to the backup server at the same time it can create a bottleneck on the backup appliance. The problem is that most backup appliances cram as many high capacity hard drives as possible into a single backup appliance. Doing so reduces the number of hard drives and the cost per GB of the system. The single set of controllers (often only one is active) has to deal with this random I/O stream and it has to do that with a limited number of hard disks to spread that I/O pattern across.

The Consolidated Full Backup Challenge

The third problem is the consolidation of the CBT backups to make a new full backup. This consolidation effort is important so that recoveries don't become too slow and data becomes vulnerable to loss. Veeam and other CBT solutions have the ability to create this full backup synthetically without having to impact the original virtual machine for yet another backup. The process requires that data from the last full and the subsequent CBT backups be essentially recovered from the backup target, brought to the Veeam backup server and merged so that an updated full image can be constructed. This full image is then stored back on the backup target.

Again, in an environment where a standard disk array is directly attached to the Veeam server, this process can happen quickly and seamlessly via the SCSI connection. But when a disk backup appliance is used a combination of problems present themselves. The first is the network connection itself. Data has to be transferred back and forth between these two connections via an un-optimized CIFS connection. The fourth problem is that the typical disk backup appliance needs to un-deduplicate, or rehydrate the data before it can be transferred across the network. The combination of hydration and network communications can make the synthetic full backup process quite slow. It is so slow in fact, that many data centers choose to forgo the apparent efficiency of a synthetic full and do a "real" full by transferring all the information from the VM to the Veeam backup server.

The Replication Challenge

The final backup problem when trying to combine VM specific backup solutions and disk based backup appliances is dealing with replication to a disaster recovery (DR) site. Both solutions have the ability to replicate data, but neither possesses an awareness that the other has that ability. Each has their own justification as to why it should be in control of the replication process.

Allowing the VM specific backup application to control and manage the replication process allows for that application to report on the completion status of the data protection process from backup to the local site to replication to the remote site. It can be aware if a replication job did not complete and attempt to retry until it does. But running the replication job on the backup server consumes server resources that may be needed to perform CBT backups, creation of synthetic fulls or managing an in-place recovery. If replication is done on the backup appliance, none of those backup server tasks are impacted, but loss of end to end protection reporting is.

The Recovery In Place Challenge

Storage Switzerland detailed the recovery-in-place problem in its webinar,"How Server Virtualization Is Breaking Disk Backup". To recap, disk backup appliances gained popularity by driving down the cost per GB of disk storage to the point that its advantages over tape based systems can be easily justified. To accomplish this, these systems leveraged very high capacity disks and data efficiency technologies like deduplication and compression. This effort has been so successful, that disk has almost replaced tape entirely as the initial backup target.

The problem is that these technologies create a problem when using a feature like recovery in place. High capacity drives means a fewer quantity of drives to service workloads and with processing intensive features like deduplication and compression in use, this can mean degraded performance when a recovery in place occurs. In short, the virtual machine that is recovered in place has to be constantly deduplicated and re-hydrated while being accessed in a production situation.

Making Disk Backup and VM Specific Backup Get Along

The typical workaround for these dilemmas is to provide exactly what Veeam and these other solutions were originally designed for - straight, unadulterated, direct attached disk. But the downside, of course, to this approach is potentially higher disk backup costs.

An alternative may be to leverage a device like ExaGrid's. It provides a native landing zone where the most active backup set is not deduplicated or compressed. This capability resolves much of the thrashing caused by deduplication during recovery in place as well as some of the performance issues related to the above backup challenges.

In its latest release, the ExaGrid solution can actually run the Veeam data mover on the backup appliance. This provides a native transfer of backup data instead of using CIFS/NFS. It also means that jobs like the consolidation of CBT backups into a new synthetic full do not need to be sent to the Veeam backup server, instead, all backups can be done on the disk backup appliance.

Conclusion

Data Centers need all the capabilities that VM specific backup is known for but they also need an economically sound data protection target. Until recently, accomplishing that objective often meant either using traditional disk arrays exclusively or using them in conjunction with a deduplication appliance. In either case, this raises the cost of the backup infrastructure and increases complexity.