WHITE PAPER

Optimizing Backup and Data Protection in Virtualized Environments

January 2009

Introduction

The promise of maximizing IT investments while minimizing complexity has resulted in widespread adoption of server virtualization. In fact, VMware users are now deploying nearly half of their new servers as virtual servers, rather than as physical servers. That said, rapid adoption of server virtualization has brought with it a variety of challenges in backup and data protection.

IT administrators have traditionally backed up their physical servers with backup agents on each server, with the backup agent installed on top of the operating system (OS) running on the physical server. Each backup agent had its own physical server on which to run, and could access all the resources belonging to that physical server without undue concern for over-utilizing the physical server.

However, this old paradigm of using a single backup application agent on each physical server can break down with large numbers of virtual machines running on a smaller set of physical servers. With the adoption of server virtualization, where multiple guest operating systems now run on a single physical server, the traditional one-backup-agent-per-OS approach presents problems. Multiple guest OS's may now require multiple backup agents to be run on a single physical server, potentially taxing the resources of the physical server and therefore the resources available to all of the virtual machines and all of the applications running on those virtual machines, on that single physical server.

Additionally, as multiple virtual machine backups with potentially overlapping backup windows are sent to one or more backup targets, landing all of those backups so that they can be protected as soon as possible is an increasing concern.

Server virtualization also brings advantages in creating second-site server backups, since all that is required for a virtual server backup is a copy of the virtual server image. However, managing all of these virtual server copies in a way such that the virtual servers can be retrieved and brought online as fast as possible can be a challenge.

Server virtualization can also increase storage utilization, as moving from physical to virtual servers often results in a larger number of combined physical and virtual servers in the environment. This can occur because the cost of generating a new virtual server is now much lower than was the cost of procuring an additional physical server. This in turn increases the amount of primary storage needed to support the additional servers, and increases the amount of secondary storage needed to back up these additional servers. As storage needs increase with greater numbers of virtual servers in the environment, the need for better visibility into the storage utilization of individual virtual machines also becomes important.

Virtual Server Backups using Disk-based Backup with Data Deduplication

While server virtualization does result in additional backup and data protection challenges, many of these challenges can be mitigated via a disk-based backup solution with data deduplication. There are a variety of these types of solutions in the market, however, each with their own approach. Therefore, when examining disk-based backup solutions using data deduplication, companies should consider the approach of each solution in the following areas:

1. *Data Deduplication Ratios.* Because multiple backups of a given virtual machine are likely to contain highly redundant data, virtual machine backups lend themselves particularly well to data deduplication. With data deduplication, virtual machine backups are compressed and de-duplicated, yielding ratios of potentially 1000s to 1, and allowing the customer tremendous savings in backup storage.

2. *Backup Job Aware Reporting*. Backup job aware reporting is the ability to view status on how much deduplication is occurring – and how much disk space is being saved – at the backup job level. This provides a number of benefits. With backup job aware reporting, the user has visibility into how well individual virtual servers are being deduplicated. It also provides the replication status of a virtual server – when the last time a given virtual server was replicated to a second site. These features give administrators the ability to better diagnose space utilization issues, since deduplication ratios for specific virtual machines are known. Network bandwidth issues may also be discovered via replication status reports. When looking at disk based backup solutions, be sure to ask about the following reporting capabilities:

    - Can the solution tell you how much storage a given backup job is using, and how well deduplication is working for that particular backup job?

    - Can the solution provide the ability to find the replication status of a given backup job – whether and when a particular backup job was replicated to a second site?

3. *Fast backups and fast restores*. Companies must consider the differing approaches of post-process vs. in-line data deduplication methods. With an in-line approach, data deduplication occurs as data is flowing into the system, and before the data is written to disk. With post-process data deduplication, the data is sent directly to disk, and data deduplication occurs after the data has landed on the disk.

Because data is being processed and deduplicated as it makes its way to disk, performing backups using an in-line approach can result in slower backups and longer backup windows. With a post-process approach, because data lands to disk first, the backup can occur at as fast a speed as your backup environment allows. The net result is high performance backups with the shortest backup windows. This means that your virtual servers can be protected faster and with the shortest possible backup window so that there is minimal intrusion into your normal IT operations.

In-line and post-processing approaches also differ in their ability to provide fast restores. With post-processing, the most recent backup is kept in its entirety, ready to be restored as quickly as possible, so that you have the ability to get an extremely fast restore time from your most recent backup. This means that the most recent version of a virtual server that has been backed up can be restored very rapidly, as it is read from disk without needing to be re-assembled from its deduplicated state. This stands in contrast to in-line approaches, where, if you needed to restore a virtual machine that had been backed up just five minutes ago, the contents of that virtual machine would need to get re-assembled through the deduplication algorithm before it could be restored, which is a much more time consuming approach. So, when looking at these solutions, be sure to ask the following regarding backup and restore times:

- How important is it to be able to back up your virtual servers as quickly as possible? Do you plan to grow the number of virtual servers that you are backing up so that you could be running up against your backup window in the future, so that the importance of faster backups might grow over time?

- How important is it to be able to restore your virtual servers as quickly as possible?

4. *Second-site virtual server protection and recovery*. The ability to replicate virtual machine copies to a second site, allowing offsite protection for virtual servers in the event of a disaster or other outage at the primary site, is another area to consider when examining disk-based backup with data deduplication approaches. It is important here to not only look at the ability to replicate the data to the second site efficiently, but to also examine the ability to quickly restore a virtual machine on the second site.

   Data deduplication makes second-site protection more efficient, as only changed data is transmitted across the WAN from the primary site to the second site. Because the amount of data that changes in a given virtual server is generally minimal, keeping a second-site copy of a virtual server up-to-date can be done easily and efficiently.

Where some approaches differ is in the area of fast recoveries on the second site. In some implementations, in order to restore the most recent copy of the virtual machine on the second site, the data making up this most recent copy must be pieced together from various parts of deduplicated data. This is potentially a costly and time consuming operation. Other approaches maintain the most recent version of the virtual machine backup in its complete form on the second site so that it can be restored as quickly as possible. Make sure to consider these vital questions:

- Do you have a need for a second site for additional protection of your virtual servers?
- How important is it to be able to recover your virtual servers at the second site as quickly as possible?

5. *Scalability.* Another important area to look at when evaluating disk-based backup with deduplication solutions is scalability. The best scaling solutions grow along with the environment in which they are working, in a manner that is less disruptive while maintaining performance characteristics over time. In some implementations, the customer installs an appliance that is sized and specified for their environment, but as the amount of data grows, the solution scales up either by adding additional only storage capacity, or by replacing the appliance with a larger, higher performance unit. This approach to scalability can create potential problems. Simply adding storage to an existing unit, for example, means that a greater amount of data is being managed by the same amount of processing, memory, and bandwidth. This can result in slower backups and longer backup windows. Replacing a lower-performing appliance with a higher-performing one, to better deal with the greater amount of data, on the other hand, is disruptive to the backup environment.

   Better approaches to scalability allow the system to maintain all the elements needed for performance as the amount of data grows. Grid-based approaches, where multiple servers that each contain processing power, memory, and bandwidth, in addition to disk, have the ability to allow growth in the amount of data to be backed up, while also packing the additional processing punch to manage this data. Grid-based approaches can also grow with much less disruption to the backup environment. Instead of replacing smaller units with larger ones, additional servers are simply added to the grid. So, when looking at the issue of scalability, be sure to consider the following:

   - How does the disk based backup with deduplication approach deal with a growing environment – will the growth of data you are backing up impact your backup window?

- Should you need to upgrade to larger systems in order to manage more data, how disruptive is that upgrade to your backup environment?

VMware Backup Methods with ExaGrid

The challenges of backing up servers in a virtualized environment opened the door for the adoption of many new backup methods over the traditional backup apps that dominate the market. Some IT administrators devised their own script-based and other home-grown methods for backing up virtual machines, and various third-party vendors emerged by productizing these types of approaches. The result is that today there are now numerous methods for performing virtual machine backups, in addition to the physical machine backup methods that existed prior to the widespread adoption of server virtualization. IT customers that have adopted server virtualization are now likely to have multiple backup procedures in place within the same data center – procedures to handle physical machine backups, and procedures to handle virtual machine backups.

Regardless of which backup methods are used, and regardless of whether a single backup method or multiple backup methods are employed, ExaGrid gives the customer a single backup system to which both physical and virtual machine backups can be targeted, while allowing the customer the flexibility to choose among the various VMware backup options that are available. Below are the VMware backup methods that are supported by ExaGrid:

- *Support for backup application agents running within virtual machines*. This method, where a backup agent is installed in each virtual server, or virtual machine (VM), is carried over from the way that physical servers are ordinarily backed up. A backup agent is installed on each virtual machine, even if there are multiple virtual machines installed on a single physical server. These backup agents are managed by the backup server, while the ExaGrid system sits behind the backup server, as a target for the backups. Virtual server and physical server backups are essentially treated the same way. This method is recommended for mission critical applications where a backup agent with application-specific optimization is needed.

- *Support for a backup agent on the VMware ESX Server Console itself*. In this method, a single backup agent is installed in the ESX Server Console. From the console, the backup agent can then perform essentially a file system backup and back up the virtual machines as single files (.vmdk files). The ExaGrid system again sits behind the backup server, as a target for the file system backup. Note that quiescing the virtual machines that are to be backed up is required in this scenario.

- *Support for backup agents running with VMware Consolidated Backup (VCB)*. VMware Consolidated Backup (VCB) is a backup application extension that facilitates VMware backups with traditional backup application agents. This method allows a backup agent to run "off-host," or off of the ESX Server itself, on what is called a "proxy server". (The proxy server currently must be a Windows 2003 server.) The backup agent uses VCB to run a script of VMware command lines, on the ESX server, to quiesce a given virtual machine to be backed up and then creates a snapshot of that virtual machine. The virtual machine snapshot is then mounted on the proxy server and made available to the backup application (typically running on the proxy server as well). The backup application then performs a backup of the virtual machine from the proxy server to the ExaGrid system (the backup target).

- *Support for direct backups of VMware virtual machines*. This method allows users who do not wish to use a traditional backup application the ability to copy virtual machines directly to the ExaGrid system. The user simply copies the .vmdk and other accessory files that correspond to a given virtual machine, directly to the ExaGrid system. This method also includes support for Vizioncore's vRanger Pro application, as well as other possible scripting solutions that directly write the appropriate virtual machine files to the ExaGrid system.

Summary

The return on investment of server virtualization is unquestioned; however the impact this has on backups must be addressed. The current range of backup and data protection issues facing companies who have moved to a virtualized environment are numerous. However with the right disk based backup approach, companies can enjoy the gains achieved with server virtualization and gain better control of the backup environment. ExaGrid's approach works extremely well with companies in this position. The ExaGrid system greatly reduces the amount of storage needed for your backups through data deduplication. ExaGrid's post-processing approach provides the fastest possible backups and restores, and extends this capability to second sites via Instant Disaster Recovery – the ability to rapidly restore your most recent backup on the second site as well. ExaGrid's backup aware reporting provides insight into how much data your backup jobs are using, how well they are being deduplicated, and how quickly they are being replicated to a second site. Finally, ExaGrid's grid-based architecture allows the system to scale easily and with minimal disruption to your environment.

About ExaGrid

ExaGrid is the leader in cost-effective and scalable disk-based backup solutions with byte-level data deduplication. A highly scalable system that works with existing backup applications, the ExaGrid system is ideal for companies looking to quickly eliminate the hassles of tape backup while reducing their existing backup windows. ExaGrid's patented approach minimizes the amount of data to be stored by providing standard data compression for the most recent backups along with byte-level data de-duplication technology for all previous backups. Customers can deploy the ExaGrid system at primary sites and secondary sites to supplement or eliminate offsite tapes with live data repositories or for disaster recovery.

ExaGrid Systems, Inc. 2000 West Park Drive Westboro, MA 01581 **1 800.868.6985** **www.exagrid.com**