

Optimizing Data Protection Operations in VMware Environments

April 2009



Data protection is critical for small and medium business (SMB) customers. Evolving business and regulatory mandates are driving ever more stringent requirements in the areas of backup window, recovery point objectives (RPO), recovery time objectives (RTO), and recovery reliability. Conventional data protection infrastructures in the SMB are still largely tape-based, and while tape is an inexpensive solution, in many cases it is not proving up to the new requirements for backup, recovery, and disaster recovery (DR). On top of that, the SMB's common lack of sophisticated administrative resources can make responding to these new challenges even more difficult, particularly for distributed environments.

Server virtualization can offer some significant advantages in resolving these data protection issues for SMBs. While the cost savings associated with server consolidation alone have been enough to justify the move to server virtualization for many companies, it's important not to overlook the opportunities it provides to optimize data protection operations. Many of these optimizations are just not available in physical server environments, and it is our recommendation that, as SMBs become comfortable with their server virtualization deployments, they should be looking at each and every one of these other optimization opportunities. For the SMB to take advantage of them, though, they must be very accessible – easy to deploy and use. In this Solution Profile, we'll identify the data protection optimizations available in the VMware vSphere environment, and show how the recent VMware Data Recovery introduction makes it very easy for SMBs to leverage them to resolve backup window, RPO, RTO, and recovery reliability concerns.

Data Protection in the SMB Today

While SMBs are being affected by many of the same issues causing data protection headaches for enterprise environments, they also face a set of unique challenges. Budgets can often provide data protection coverage for only the most mission-critical applications, and sophisticated data protection administrative expertise can be in very short supply. Existing tape-based backup

infrastructures are often unable to address backup window, RPO, RTO, and recovery reliability issues which are top of mind for most data protection administrators. And when evaluating new solutions, it's clear that SMBs have a strong preference for solutions that are easy to deploy and manage.

Because of the management advantages and rapid return on investment (ROI) that server virtualization provides, it has been very popular with SMB customers. As customers

S O L U T I O N P R O F I L E

move to these environments, they often migrate their data protection processes to them unchanged. When evolving computing environments, it's smart to use a phased approach to making changes, but in our opinion you're making a mistake if you're not also looking at modifying your data protection processes soon after completing the transition to server virtualization. Because of the way server virtualization works, data protection processes designed for physical server environments can perform poorly. But take heart: there are data protection optimizations available in virtual server environments that can not only restore that performance, but go far beyond anything achievable in physical server environments.

Physical Data Protection Processes

In physical server environments in the SMB, applications are generally paired with servers on a 1 to 1 basis. As a result, most of these servers are woefully underutilized, but administrators are hesitant to run them at capacity because of the risks this poses in dealing with infrequent spikes in CPU and memory usage. When servers are virtualized, multiple servers are effectively consolidated onto a single physical server platform, and much of the initial cost savings with server virtualization accrues as a result of this.

For most SMBs, data protection requirements demanded that each physical server host a backup agent. Data protection operations include activities, such as backup and restore, that tend to put a high load on physical servers, particularly in the I/O area. Underutilized physical servers generally have enough excess I/O capacity in them to meet the data transfer requirements during

backup, but it is very expensive to keep this capacity around unused most of the time. Virtualized servers, on the other hand, are running on physical servers and exhibit much higher average utilization rates, but if multiple virtual servers on a single physical server are being backed up at the same time, I/O can become a bottleneck which slows backup. This is what can happen if data protection processes designed for physical servers are just transferred over to virtual server environments without modification.

Awareness

In our experience, many of the SMBs that have deployed server virtualization do not fully understand the data protection optimizations available to them in this new environment. They likely have experienced the effects of moving backup and restore processes over from their physical environments unchanged, but are unsure how they should evaluate other options.

Ease of Use Considerations

Because of the lack of sophisticated administrative expertise often found in smaller environments, data protection optimizations must not just be available – they must also be easy to use. This is a factor that, along with the awareness issue, has kept many SMBs from considering and deploying such solutions.

Data Protection Optimizations

Virtual machine (VM) environments offer the opportunity to optimize data protection operations to shorten or remove backup windows, significantly improve RPO/RTO, and improve recovery reliability relative to

S O L U T I O N P R O F I L E

the conventional tape-based operations in use in most SMBs. The best way to leverage these optimizations requires that companies move to disk-based data protection operations built around centralized network storage architectures. Disk-based data protection offers significant improvements in physical machine environments as well, but certain features of the VMware environment allow data protection to be optimized in ways not possible in physical server environments. Centralized storage architectures like SAN enable the use of interesting technologies for the migration of servers and storage which offer important advantages when optimizing data protection operations. This brings meaningful improvements for both local and remote data protection activities.

When used for backup, disk performs better than tape for most backup operations that are performed on a regular basis, like daily backups and file-level restores. The performance characteristics of disk result in shorter backup windows and more reliable recoveries as well.

Storage is often consolidated onto a network during server consolidation projects, and centralized storage platforms, when combined with the VMware vStorage VMFS clustered file system, provide a strong foundation for optimized data protection operations. Administrators are not required to move to centralized storage during server virtualization projects, but Taneja Group would highly recommend it because of the benefits it can provide for managing data protection, disaster recovery, application availability, and balancing resource utilization among virtual servers (benefits that go

well beyond just data protection concerns). In discussing data protection optimizations achievable in VMware environments, we will assume that administrators have deployed disk-based data protection, a capability supported in almost all major backup software products, and centralized, networked (as opposed to direct attached) storage. Moving to disk-based data protection is critical because the use of disk opens up access to disk-based technologies, such as snapshots, data de-duplication, and replication, that offer real advantages for cost-effectively managing business continuity requirements (such as data protection and disaster recovery) to the highest levels.

For appropriately configured environments, the vSphere platform offers local data protection optimizations in three areas: snapshot backups, data de-duplication, and restore operations. We'll address each of these three areas independently.

Snapshot Backups

Snapshot backups can be used to effectively remove backup windows, allowing backup operations to occur regularly with *no impact* on production applications. Dumping data directly to tape during conventional backup operations can be a lengthy process limited by the performance characteristics of tape. Recoverable images on disk, however, can be almost instantly created with snapshot products. A separate server, referred to as a "proxy" server, can then take over the ownership of the snapshot and back the data up in a manner completely de-coupled from the server being "backed up". There is significant value in being able to perform backups with no impact to production.

S O L U T I O N P R O F I L E

Historically, implementing snapshot backup operations has required some custom development by the administrator, increasing the complexity of using these approaches. Vendors have begun to offer better approaches for the use of snapshot backup that leverage standardized APIs, and do not require custom scripting. For Windows, Microsoft has introduced Volume Shadow-copy Services (VSS), an interface that allows recoverable snapshots to be created without disrupting applications. Because VSS is open, third party vendors like VMware can take advantage of it to build off-the-shelf, productized solutions that do not require any custom development, are easy to deploy, and are fully supported.

Data De-Duplication

Data de-duplication is used to reduce the amount of physical storage capacity required to store a given amount of data. Data de-duplication segments data streams into “chunks”, fingerprints the chunks, and then stores the chunks in an index. Redundant chunks are identified by the use of fingerprint comparisons (incoming new data vs data the system has already seen) and removed and replaced with pointers. The removal of duplicate chunks is what results in the savings in storage capacity. When de-duplicated data is read back, checksums are used to validate that the “reconstituted data” exactly matches the data that was originally de-duplicated.

Data de-duplication can be deployed either at the source (the server where data is created) or at a storage target (the storage where the data is stored). The advantage to source-based de-duplication is that backup data is

capacity optimized before it is sent across the wide area network (WAN), and sending a lot less data across the WAN can shrink backup timeframes significantly. Conventional source-based de-duplication consumes backup client resources though, which can impact application performance if backups are being done on-line. Target-based de-duplication puts no impact on backup client resources, but data is not capacity optimized until it hits the storage target, which means that there are no time and bandwidth savings as the backup is sent across the WAN. By combining data de-duplication with snapshot backups though, administrators can get the space-saving advantages of de-duplication without impacting production applications during backup.

Data de-duplication provides immediate savings to companies by reducing storage infrastructure requirements, saving on power, cooling, floorspace, management overhead, and helping to defer new storage purchases even in high growth data environments. As a technology, data de-duplication is approaching mainstream status, and it is available from a number of vendors either integrated with backup software (source-based data de-duplication) or as an appliance (which acts as a storage target).

Backup data presents lots of opportunities for de-duplication. Backup data does not change much day to day, yet backups must be done regularly to ensure the ability to quickly recover to current data. This means that a large percentage of data in consecutive daily backups is redundant. Our experience indicates that, with consistent use of data de-duplication technologies over time, most

S O L U T I O N P R O F I L E

data protection administrators can reduce the amount of raw physical capacity required to store backup data by 90% - 95%. But storage infrastructure cost savings are not the only benefits of using de-duplication technology – the much lower volumes of stored data can result in much shorter backup windows (if snapshot backup is not being used), faster restores (because less data is being moved by each activity), and significant savings in bandwidth when de-duplicated data is replicated to a remote site for DR purposes.

Restore Operations

Most restore requests are for one or more files. Very few restores require all the data in a given system to be restored. Having the ability to select and restore individual files directly can save a lot of time and effort for backup administrators, although the ability to restore entire systems must exist as well to address the rare event when it is needed. Ideally what is desired is the ability to select the backup, search and identify the desired file or files by name, and then perform the restore directly to the requestor. When this sort of “file-level restore” can be safely accomplished directly by the end user, without having to involve a backup administrator, it reduces the general administrative burden even more.

Traditional backup operations have focused on data files, but when entire systems must be restored (due to catastrophic failures) additional data is required as well. Re-creating a system from scratch includes installing an operating system and one or more applications, as well as configuring the system for the particular environment.

Operating system patches and custom configuration tweaks may also have to be applied. When all of this work must be done manually to prepare a new system before data is restored to it, it can take hours or days per system, depending on the size of the system. The basic problem is that all of this additional data (operating system, system configuration data, application binaries, patches, and configuration tweaks), referred to as “system data”, cannot just be “backed up” like data files can; for physical systems, they require separate tools specially designed to back this information up and then potentially significant operator involvement to re-assemble this data, along with the data files, to re-create the system. Tools in use today to capture this system data include native operating system utilities, products like Symantec Ghost (for smaller systems), or various bare metal restore products from vendors like Acronis, Cristi, and Unitrends that use disk-based imaging technology.

Server virtualization technologies like VMware vSphere change this picture considerably. Entire VM images, including both system and data files, are encapsulated in VMDK files, and these files can be backed up using standard file-level backup products and tools. This encapsulated image of the entire VM is referred to as a “bootable” image, meaning that a new system that exactly matches the old one can be directly started from this image (without having to go through a manual rebuild process). This offers two key advantages that physical server environments cannot match:

- Standard file-based backup products can be used to create and maintain current

S O L U T I O N P R O F I L E

bootable images of any VM, and that bootable image can include all of the information required to completely restore an entire system, making server-level restores in VM environments much faster and easier (with far fewer steps) than traditional server-level restores in physical machine environments

- Server migrations and the creation of new servers based on established templates (e.g. the “golden” image for all Windows file servers) are a variant of this optimized server-level restore, and just as fast and easy

Server-level restores, fortunately, tend to be rare events, but creating new servers is a frequent event, and the flexibility to perform server migrations almost instantaneously (and without impact to running applications, depending on which tools you’re using) offers a lot of value for high availability and maintenance operations.

VMware Data Recovery

The introduction of data de-duplication technologies is new for VMware (although it is a very familiar technology in the industry), but the other data protection optimizations discussed have already been available to vSphere customers. What is new with VMware Data Recovery, however, is that VMware has packaged this functionality into a virtual appliance that is much easier for SMBs in particular to deploy and manage. Virtual appliances are cost-effective because they allow the deployment of dedicated functionality without the expense of adding another physical server. The Data Recovery virtual appliance is pre-configured to provide

the data protection optimizations discussed earlier in the simplest possible manner, making the power of vSphere extremely accessible to smaller customers who may have less sophisticated and/or less extensive data protection expertise.

Snapshot Backup Operations

For Windows environments, Data Recovery uses the VMware vStorage APIs for Data Protection (formerly VMware Consolidated Backup (VCB)), a backup framework included with vSphere at no additional charge, and VSS to regularly create recoverable snapshot backups, per schedules established by the administrator. Snapshots are created without impacting Windows application performance using vStorage APIs, and ownership of those snapshots is then assumed by the virtual appliance using a new “hot add” feature. Because the underlying storage is network-attached, this action is performed without requiring any data movement. This same approach can be used against other guest operating system environments besides Windows (e.g. Linux, Solaris, etc.) but will require some application-specific scripting to obtain recoverable snapshots as those operating systems lack a comprehensive snapshot utility like VSS.

vStorage APIs coordinate snapshot creation at the hypervisor level, an approach that is much more efficient and has much less impact on production servers than approaches which operate at the guest operating system level. When vStorage APIs are used to perform backups at the hypervisor level, backup agents from third party vendors are not required.

S O L U T I O N P R O F I L E

Data De-Duplication

Once the virtual appliance owns the snapshot, it can then apply VMware's data de-duplication algorithms to reduce the amount of physical storage capacity the backup requires. The virtual appliance hosts the de-duplication index, and its central location in the virtual appliance tends to result in higher de-duplication ratios than if each production VM was performing its own de-duplication work independently. What's interesting about this approach is not only the ease with which it is deployed and managed, but also its efficiency. In a virtualized server environment, other source-based data de-duplication products that are integrated into a backup client draw from VM resources to perform their work, impacting the performance of VMs being backed up, sometimes considerably. VMware's approach moves all de-duplication work to the dedicated appliance, completely offloading the VMs being backed up, and completes it before the data is ever sent out across a network (such as would occur as backup software is used to back up the snapshot backup). And this approach is very easy to use: administrators just enable data de-duplication, and Data Recovery will de-duplicate all VMDK volumes that it knows about. Policies for snapshot backup can be managed from the virtual appliance for any of these de-duplicated volumes.

VMware's approach to data de-duplication is somewhat of a hybrid between source- and target-based approaches. Data de-duplication occurs at the source so that the amount of data sent across networks during backup operations is kept to a minimum, but it does it in such a way that does not impact

production server (VM) resources. This approach is unique in the industry. Any concerns about ensuring that the virtual appliance has access to the resources it needs without impacting the VMs it is backing up can be handled with VMware's Distributed Resource Scheduler (DRS), a component of the vSphere platform.

Recovery Operations

Data Recovery can be configured to support user-administered file-level restores. Access controls are established by the backup administrator so that end users can safely perform their own recovery operations without any risk to data that is not their own.

Because VMDK files encapsulate all data (system and file level), the benefits of snapshot backups and data de-duplication apply to the creation and maintenance of bootable images that enable fast, easy server-level recoveries and migrations or the creation of new VMs. VMware Data Recovery captures all of this information as part of the standard backup process, ensuring that administrators always have an up-to-date image to use for recovery and/or migration purposes. There is no doubt that this is a much easier and more efficient approach than that commonly in use in physical machine environments today that requires separate products and schedules to back up system configuration and data files.

Centralized Management

From a management point of view, Data Recovery is well integrated into vSphere, making management operations easier. Data Recovery operations are all managed from within the centralized vSphere console,

S O L U T I O N P R O F I L E

providing efficient management operations from a single pane of glass. Security fits within the vSphere framework, so that Data Recovery access is controlled through the vSphere single log on. Other vSphere components are aware of Data Recovery operations so that the integrated management of all components is optimized. One example of this is that if backups are in process when VMs are re-located using VMotion, those backups will still complete without requiring operator intervention.

Taneja Group Opinion

For vSphere customers, there is a great set of tools designed to enable data protection optimizations that are just not achievable in physical machine environments. These tools include vStorage APIs, DRS, and now their data de-duplication technology, and make it possible for customers to use a common data protection solution that covers all operating environments and applications within a server virtualization platform. VMware Data Recovery makes it easy for SMBs to take advantage of these tools by packaging a comprehensive data recovery solution, pre-configured for use in VMware environments, into a virtual appliance that is easy to deploy and manage. Specifically, here's how VMware Data Recovery makes things easier:

- Install a single, pre-configured appliance (that doesn't require additional hardware) that handles snapshot backup, data de-duplication, and server-level recovery

- Leverage the latest vStorage API data protection optimizations for VM environments
- Dispense with the need for third party backup agents on all VMs that need to be backed up, saving on license and maintenance costs
- Perform data protection operations without having to impact the availability of Windows applications
- Lower backup infrastructure and network bandwidth costs through the use of data de-duplication technologies without impacting production server performance
- Use the same data protection processes for all operating systems and applications in your virtualized server environment

VMware has had a rich tool set available to more sophisticated users in the past that streamline data protection operations in virtual server environments, and their Data Recovery product introduction now makes those optimizations much more accessible to SMB customers. If you are a vSphere customer today or are considering virtualization technology and have concerns about backup windows, RPO/RTO, and recovery reliability, you need to consider using the data protection optimizations available to you. As an SMB, the easiest way to take advantage of them is through Data Recovery. We recommend that all SMBs using VMware vSphere evaluate how VMware Data Recovery can improve data protection operations in their environments.



S O L U T I O N P R O F I L E

***NOTICE:** The information and product recommendations made by the TANEJA GROUP are based upon public information and sources and may also include personal opinions both of the TANEJA GROUP and others, all of which we believe to be accurate and reliable. However, as market conditions change and not within our control, the information and recommendations are made without warranty of any kind. All product names used and mentioned herein are the trademarks of their respective owners. The TANEJA GROUP, Inc. assumes no responsibility or liability for any damages whatsoever (including incidental, consequential or otherwise), caused by your use of, or reliance upon, the information and recommendations presented herein, nor for any inadvertent errors which may appear in this document.*