

WHITE PAPER



Virtualization, Deduplication and the Data Protection Connection

Reinventing IT Infrastructure Consolidation
with Efficient Data Protection

Virtualization, Deduplication, and the Data Protection Connection

Reinventing IT Infrastructure Consolidation with Efficient Data Protection

Introduction	3
The Data Protection Connection	3
What is Efficient Data Protection?	4
How Efficient Data Protection Delivers Data Center Efficiencies	4
Considerations and Concerns	5
How is this model implemented?	6
How safe are the backup images?	6
Is redundant data deduplicated across enterprise or only across server?	6
Is this model scalable?	6
What are the costs?	6
Looking Toward the Future	7

Introduction

As data center resource consolidation gains momentum, many organizations are missing the big picture. For reducing primary server sprawl, they are looking at virtualization solutions in isolation; and for cutting down on secondary backup storage, they are focusing on deduplication solutions in isolation. This narrowly focused approach neglects consideration of the full landscape.

When viewing the full landscape of the data center, one notices that the two growing consolidation trends, virtualization and deduplication, converge. This convergence points to a striking and ironic connection: inefficiencies in the traditional backup and restore model. Data protection is the nexus that connects the two trends, and gaining a handle on data protection is fundamental to data center efficiency and future data center growth.

If data protection were inherently efficient, today's rigorous and sometimes frantic demand for data reduction on the secondary side, and the backup contention menace that unbalances so many virtualization experiences on the primary side, could both be eliminated or at least be brought under control. In other words, efficient data protection is the sauce that can sweeten both your data storage issues and your server sprawl challenges in a single spoonful, while at the same time adding

performance enhancements, hardware savings, improved protection, and administrative ease-of-use.

The Data Protection Connection

Stated simply, inefficient backup both underlies the need for deduplication and imposes debilitating limitations on virtualization.

Deduplication solutions exist solely because of inefficiencies in traditional data protection. That is, traditional backup results in multiple copies of the entire IT environment on secondary storage. Explosive data growth has made those multiples larger than ever, and the need for faster backup performance to accommodate more data has necessitated the move from tape backup to more expensive disk backup. The result is that secondary disk data reduction has become an unwished for necessity.

Furthermore, as many organizations that have attempted virtualization know, server virtualization projects are routinely jeopardized by inefficiencies in traditional backup and restore.

At the root is the fact that in a virtualized environment, numerous virtual machines (VMs) share the resources of the single physical VM host. Disk resources, host CPU time, and network bandwidth are particularly limited. On a large virtual machine host with many VMs, competing backup

KEY FINDINGS

- Current data protection solutions for Data Center Consolidation are limited in scope and scalability.
- Efficient data protection benefits surpass data deduplication 'only' solutions.
- Efficient data protection alleviates the resource contention problems that plague server virtualization implementations.
- Efficient data protection enhances system performance, reduces hardware purchases, and improves data and application protection.

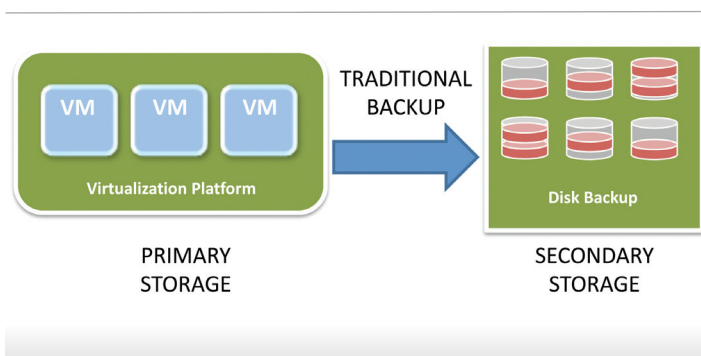
KEY PERFORMANCE INDICATORS

- **Low impact to system and network** during backup, enhancing value of virtualization investment.
- **Improved backup speed and performance** in virtual environment, enabling frequent recovery points.
- **Low secondary storage requirements** and disk purchases, eliminating need for data deduplication solution purchases.
- **Ability to recover either a file or system very quickly.**
- **Administrative ease of use.**

jobs have been known to bring the host to a grinding halt – the result is that the backup jobs fail or hang, leaving critical data unprotected. Resource-intensive backups also negatively impact the performance and response time of applications running on other VMs on the same host.

Some organizations respond to this by staggering VM backups and running only a few at a time, but this greatly extends and often exceeds desired backup windows. Others take the even more drastic step of shutting VMs down in order to back them up, an unacceptable process in today's 24x7 information technology environments, and a method that greatly limits the types of applications that can be virtualized.

Figure 1:
Traditional Disk to Disk Backup in a Virtual Environment



The limitations of these traditional approaches to data center consolidation can be negated through efficient data protection.

What is Efficient Data Protection?

An understanding of the big-picture view will help crystallize the fact that efficient data protection alleviates pressure points throughout an organization's entire IT infrastructure and data protection process, and successfully positions the organization for future growth.

In practice, very few solutions in the marketplace offer truly efficient data protection, primarily because of its technical complexity.

IN THEORY, EFFICIENT DATA PROTECTION IS SIMPLE

- Very low-impact server-level backup snapshots without complex hashing
- Small, frequent data transfers
- Storage of data on secondary disk only once
- Intelligent storage on secondary disk for fast recovery of any recovery point

Efficient data protection is achieved by deftly tracking changed blocks on the backup client below the file system. By doing so, no strain is placed on the resident applications, open files are not an issue, and the file system is not impacted. Once the blocks are identified, only new blocks or blocks that have changed since the last backup are transferred to the backup target. Note that truly efficient data protection is not achieved through recently-introduced "source side deduplication" technology, which actually increases the performance degradation on backup clients as well as extending backup time.

After the data has been transferred, the efficient data protection model must exploit an intricate system of indexes and pointers so that server-level backups, although tiny and consisting only of changed blocks, are made to appear on disk storage as full, immediately recoverable, backup images. Put another way, each set of changed blocks is immediately assimilated with previous server backups to form full backup images, though the amount of additional storage consumed is no more than the few blocks that were transferred. Prior point-in-time backups also remain immediately recoverable.

How Efficient Data Protection Delivers Data Center Efficiencies

If "efficient data protection," as defined in the section above, could be implemented easily, it would seem to provide an effective, integrated approach to data center consolidation – addressing sprawl not only on primary and

“Simply stated, inefficient backup both underlies the need for deduplication and imposes debilitating limitations on virtualization.”

secondary storage but throughout the process and network. At least five benefits are immediately achieved with this data protection model:

- Low system impact. Efficient data protection eliminates the contention for VM host resources by alleviating CPU and I/O impact at backup.

Reduction of network bottlenecks. The fact that the backup snapshots themselves are so small eliminates network bottlenecks as the backup image travels from VM to LAN to SAN to backup target disk.

- Data and storage reduction. The data coming over is already small and non-duplicative, so additional data reduction on secondary disk storage is unneeded. Because only new blocks and changed blocks are sent to and stored on secondary disk, no multiples of the IT environment are saved on the targets.
- Fast Backup. Because the backup snapshots and transfers are fast, frequent backups can be performed making numerous recovery points available throughout each day.
- Fast Recovery. The ability to reassemble the backup images very quickly at the destination allows an organization to meet recovery time objectives. Further, though the backup process occurs at an efficient block-level, restores can be managed with file level granularity.

Additionally, the efficient data protection model ideally delivers administrative ease of use by providing centralized scheduling, reporting, and maintenance, elimination of complex proxy server setups, and unified platform support.

Considerations and Concerns

Although efficient data protection sounds like an ideal data protection and data center consolidation model, it raises many practical questions including:

- How is this model implemented?
- How safe are the backup images?
- Is redundant data deduplicated across enterprise or only across server?
- Is this model scalable?
- What are the costs?

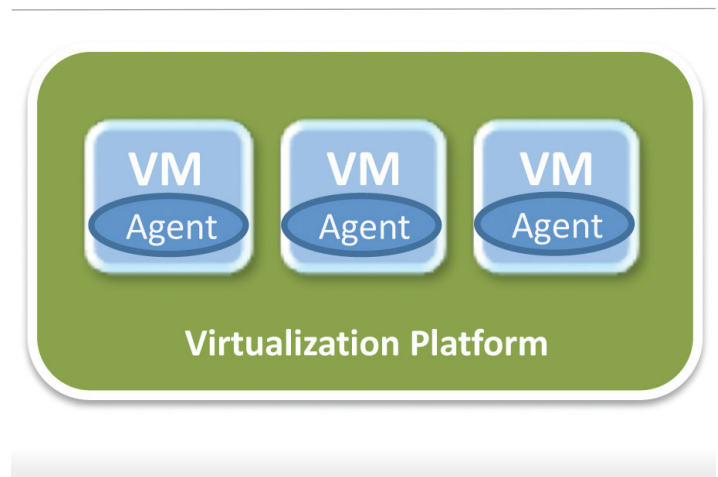
How is this model implemented?

To implement, the efficient data protection agent would be installed on each virtual server/machine or physical server that needs to be backed up. In virtualized IT environments, this type of setup is commonly referred to as a guest backup methodology. The great advantage in this methodology is that a data protection administrator uses only one product and process for physical, virtual, mixed, or migrating environments – both on the backup and recovery side. During consolidation, as physical resources are virtualized, the data protection element is essentially seamless with the guest backup methodology.

Because this model delivers the same data protection solution for both physical and virtual servers, including application servers and even desktops, administrative efficiencies are achieved in terms of training, support and reporting.

With regard to guest backup, concerns are often expressed about the fact that a backup agent needs to be added to every new virtual machine. This concern is generally overstated because each VM needs to be provisioned anyway with an operating system and other commonly deployed applications and software. New virtual machines cloned from a base system will already include the data protection agent.

Figure 2:
Guest Backup Methodology



How safe are the backup images?

Your secondary disks are by their very nature exposed to loss for many reasons including hardware malfunction, human error, and natural disaster. Such exposure can be reduced by utilizing high-grade disk and RAID configurations. Because the efficient data protection model allows you to use any disk type or disk manufacturer as a target disk, for cost-efficiency you may choose to back up vital servers to more secure disk and less vital servers to less secure disk. Nevertheless, with this model, tertiary backup would seem desirable, either through disk-to-tape archiving or disk-to-disk replication.

Is redundant data deduplicated across enterprise or only across server?

There is certainly a distinction between enterprise level deduplication and the type of data reduction performed by this model. Enterprise level deduplication eliminates redundant data not only on an individual server but across servers. So if an exact copy of a file or database or (in the case of some products) a set of blocks appears on a sister machine in your enterprise, it is stored only once on secondary storage. Enterprise deduplication does in fact achieve slightly more secondary disk data reduction than the efficient data protection model would because it works across and within servers whereas the efficient data protection model works only within servers. In the unlikely event that it becomes necessary to achieve that marginal additional data reduction, one could consider a hybrid approach combining the efficient data protection model with a target-side deduplication solution. This achieves the full benefits of enterprise deduplication without losing the other benefits – light system impact at source, elimination of network bottlenecks at backup, frequent recovery points, and so forth. The efficient data protection model should interoperate well with any target side deduplication system.

Is this model scalable?

The efficient data protection model is inherently scalable because very little additional secondary storage space is required for each successive backup. On the primary side, as backup source machines – either physical or virtual

– are added, one simply needs to install a backup agent on that machine. One concern is that the backup catalog could become large and unwieldy, so catalog efficiency is definitely a consideration when selecting an efficient data protection solution.

The question of scalability also highlights deficiencies in the commonly deployed data center consolidation solutions in use today. On the primary side, many VM users, to bypass the resource contention issue, typically reconfigure backup schedules so they don't conflict. However, it doesn't take long to realize that as data volumes grow and backup windows shrink, this solution soon dissolves into an unworkable snarl. A popular alternative has been to use a proxy server to take the strain off the host, but this alternative has been proven

to be overly complex and slows recovery. On the target side, any traditional deduplication solution, unless fortified by efficient data protection, loses efficiency over time. This results from the fact that each incoming data stream needs to be compared with an ever-growing history of previously stored data.

What are the costs?

Not surprisingly, replacing your data protection solution has a cost in terms of both money and effort. But, as has been shown, traditional backup and restore is counter-productive when implemented in virtual environments and it incurs a heavy storage cost in any disk-to-disk backup environment. Many organizations are discovering this unfortunate fact too late, and as a result they are running with different data protection products for physical versus virtual environments. A 2009 industry report indicates that two-thirds of all organizations have recently replaced all or part of their data protection solution or have expressed an interest in doing so in the next year.

To ascertain the value, or ROI, of an efficient data protection model, you need to compare the costs with all the offsetting benefits. Since you already incur software license fees and maintenance fees for your existing traditional data protection solution, your new costs are implementation and training. The benefits are less tangible but fairly easy to enumerate.

“Efficient backup will alleviate stresses throughout your entire IT infrastructure, enhance performance, improve protection, reduce hardware purchases, and successfully position you for future growth.”

They can be broken down into three categories: hardware utilization, ease of use, and performance (see box).

When compared with these vital benefits, the cost of implementing an efficient data protection solution seems negligible. And when you think about the costs saved, one can infer that investing in efficient data protection saves money.

Looking Toward the Future

Server and data sprawl, increasingly stringent application demands, and other artifacts of the twenty-first century business culture have made data center consolidation the number one priority of IT departments worldwide. Yet current approaches to data center consolidation are limited in scope as well as scalability. The period when your organization is making consolidation decisions is an excellent time to consider upgrading your traditional backup solution to one that employs the efficient data protection paradigm. This will alleviate stresses throughout your entire IT infrastructure, enhance performance, improve protection, reduce hardware purchases, and successfully position you for future growth.

HARDWARE AND EASE-OF-USE RELATED BENEFITS

- Reduce secondary storage requirements and disk purchases
- Ability to add more VMs on a single host
- Eliminate need to purchase expensive deduplication solution
- Ability to use same data protection model for physical, virtual, or mixed environments.
- Scalability and positioning for future growth.

PERFORMANCE RELATED BENEFITS

- Significantly improved backup performance in virtual environment
- More frequent recovery points
- Ability to recover either a file or system very quickly
- Minimal impact to network performance during backup
- Minimal impact to system performance (CPU) during backup

About Syncsort

Syncsort is a global software company that helps the world's most successful organizations rethink the economics of data. Syncsort provides extreme data performance and rapid time to value through easy to use data integration and data protection solutions. With over 12,000 deployments, Syncsort has transformed decision making and delivered more profitable results to thousands of customers worldwide. For more information visit: www.syncsort.com

syncsort RETHINK THE ECONOMICS OF DATA[®]

50 Tice Boulevard, Woodcliff Lake, NJ 07677

201.930.8200 | www.syncsort.com