White Paper

Why StoreOnce *Federated Deduplication* Matters to HP—and Should to You, Too

By Jason Buffington, Senior Analyst

June 2014

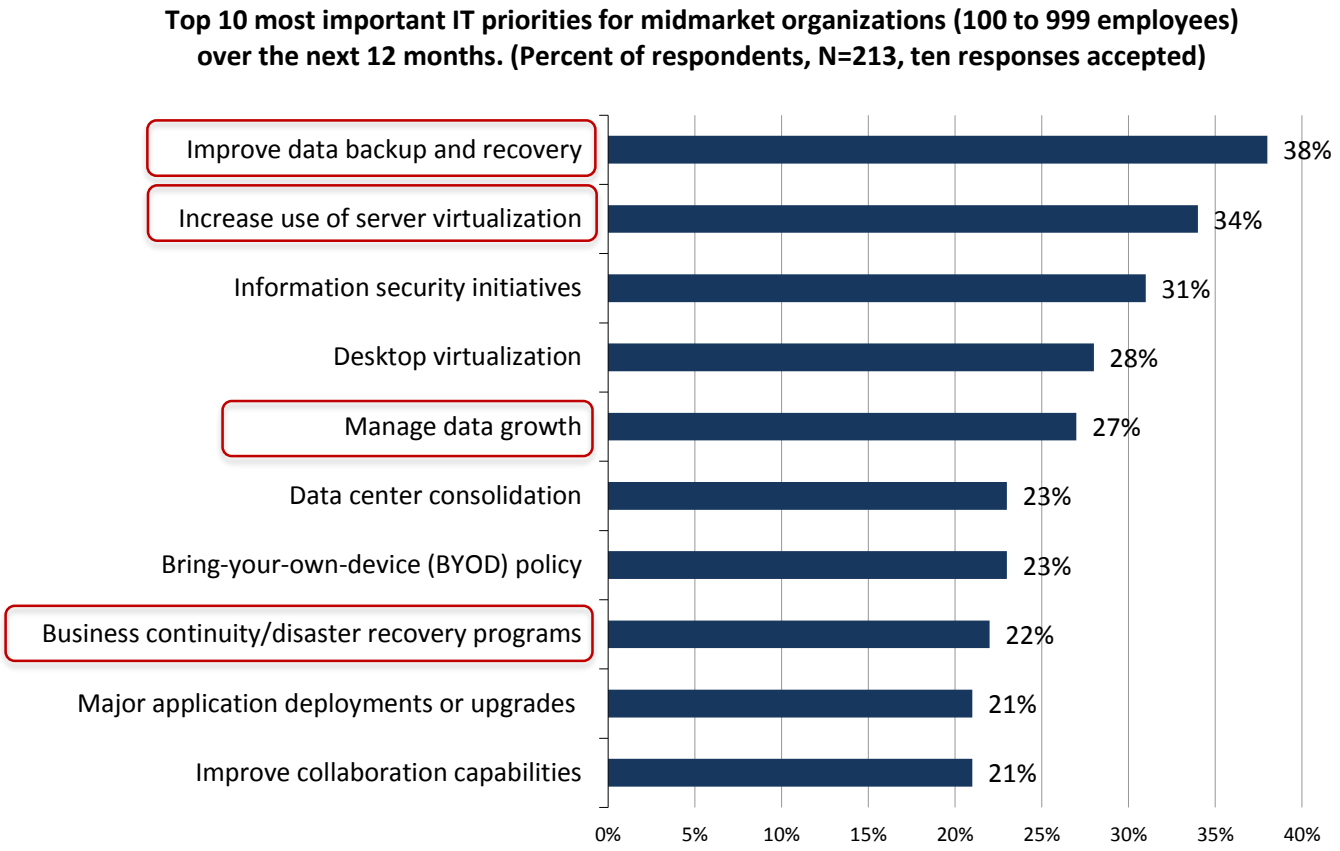
This ESG White Paper was commissioned by HP and is distributed under license from ESG.

Contents

Introduction

According to ESG's *2014 IT Spending Intentions Survey*,[1] improving data backup and recovery is the most often mentioned priority for midmarket organizations for the third year in a row (see Figure 1). It is the third most frequently cited priority by enterprises and midmarket organizations overall. Additionally, business continuity and disaster recovery (BC/DR) also appear in the top-ten priority lists of both midmarket and enterprise respondents.

Figure 1. Most Important IT Priorities for Midmarket Organizations in 2014



Source: Enterprise Strategy Group, 2014.

Data protection is usually in the top ten, but why does it continue to be the top IT initiative for midsize organizations specifically and near the top of the list overall? To understand, one only has to look at the other priorities in the list.

Server virtualization: As virtualization becomes more prevalent, VM sprawl consumes more production storage, which in turn exacerbates legacy approaches to disk-based protection. In addition, not all backup technologies adequately protect virtual machines. They often require integration with backup APIs such as VMware vStorage VADP or Microsoft Volume Shadow Copy Services (VSS), which not every backup solution has chosen to implement or support to the same degree. In fact, every challenge related to protecting virtualized environments is even more daunting in a private cloud architecture, where self-service portals and elastic load monitoring create new virtualized resources dynamically without any IT interaction with—much less awareness or automation of—backup processes.

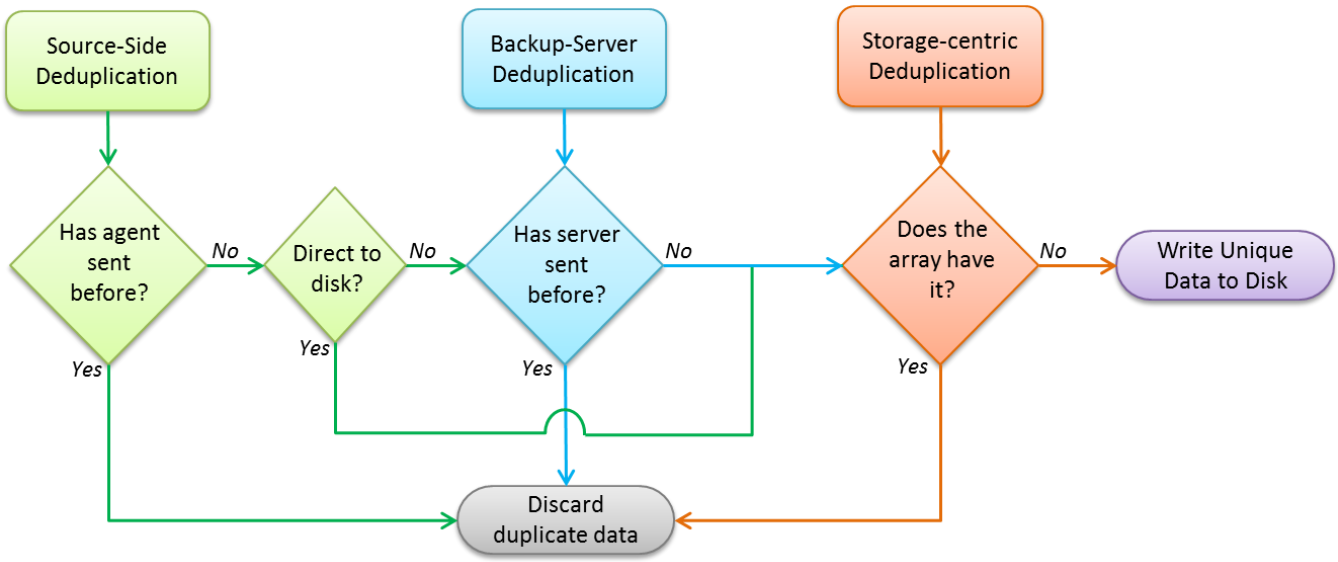
Data growth: Along with VM sprawl, every other workload continues to grow. Databases grow into "big data," e-mail mailboxes continue to expand (with retention limiting the use of offline folders), and unstructured data grows exorbitantly, often with a huge amount of undiscovered redundancy. And of course, as production data grows, so do disk-based backup stores (often at three to five times the production size).

[1] Source: ESG Research Report, *2014 IT Spending Intentions Survey*, February 2014.

Deduplication Today

With so many major IT challenges related to or caused by storage demands in production data or the backup system that protects it, deduplication has never been more of a priority or necessity. That being said, deduplication is no longer relegated to simply being part of "smarter storage" but can instead be seen throughout the data protection process (see Figure 2).

Figure 2. The Good, Better, and Best of Deduplication



Source: Enterprise Strategy Group, 2014.

Deduplication technologies provide organizations with substantial data storage efficiency advantages when deployed as a part of their data protection infrastructures. Although dedupe does not stop the growth of protected data, it does provide an effective strategy for controlling that growth (at least within one's backups).

Deduplication technologies and methodologies continue to evolve as vendors develop improved ways of ensuring that data is transported and stored efficiently. The algorithms that analyze stored data are constantly improving in efficiency, with even a fraction of a percent improvement leading to gigabytes of saved storage. In addition to the algorithm improvements, the "location" where deduplication occurs is also evolving: from simply being enhancements within secondary storage arrays, to working across the network within the data path and across multiple devices, employing a "*Good, Better, Best*" approach to optimizing deduplication.

Dedupe 1.0—Optimized Storage Is "Good"

Early deduplication appliances, while revolutionary in their day, performed deduplication strictly within the disk array itself. In the majority of "Dedupe 1.0" scenarios, the backup server treated the storage device like it would any other storage device and wrote all of its data to it. Even when the deduplicated storage array already held a significant number of items that the backup server was again transmitting, the backup server was unaware of it. Those redundant elements were discarded upon receipt by the deduplication appliance, as shown in the orange (right-side) portion of Figure 2.

From a data protection perspective, although organizations did reduce their costs by storing backup data more efficiently, the time required to perform a backup wasn't improved because the same amount of backup data still needed to be written to the Dedupe 1.0 backup device. And of course, if the backup server needed to replicate the data anywhere else, the data would be fully read and transmitted from one backup server to the next, and then written in full to another (possibly deduplication-capable) storage device at the alternate location.

Dedupe 1.5—Smarter Backup Servers Are "Better"

Dedupe 1.5 improves upon Dedupe 1.0 by enabling the backup server to only send data to the storage device that isn't *already* present on that device. In effect, the backup server becomes smarter—aware of what is in the deduplicated storage device. Dedupe 1.5 can also improve the speed at which backups happen because the backup server may no longer be the bottleneck in sending data (as less information is actually sent to the backup storage device). Dedicated backup appliances that utilize deduplication are often thought of as Dedupe 1.5 because the backup software and deduplication storage are within the same device.

Certainly better than Dedupe 1.0, Dedupe 1.5 does not typically solve the challenge of replicating data between sites without "rehydrating." The backup server will read and send the whole data set from the deduplicated storage to another offsite backup server and its deduplicated storage regardless of what the secondary deduplication device might already have.

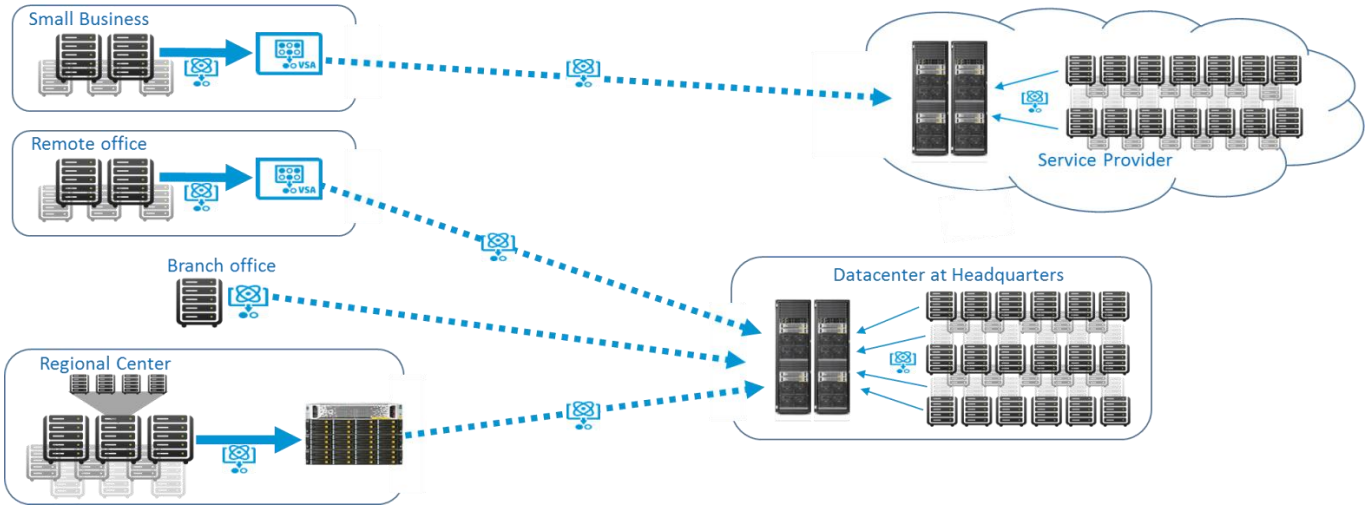
Dedupe 2.0—Client-side Deduplication Is "Best"

Dedupe 2.0 leverages intelligence and awareness at the source, backup server, *and* storage device. The awareness of what data is already in the deduplicated storage and the decision to send new data (or not to) occur within the production server instead of the backup server or deduplicated storage. As a result, network-related savings begin at the production server—backups are often significantly faster because only changed data is transmitted from the production server to the storage solution.

Client-side deduplication can be especially effective in two common scenarios:

- **When the production environment is particularly susceptible to duplication, such as in highly virtualized environments** in which new VMs are constantly spawning, often from the same base OS and application image. In these cases, by deduplicating within the host, a significant amount of redundancy across VMs is eliminated from the backup stream before ever leaving the host.
- **Deduplication is key in ROBO scenarios** because even the backup server is often at the centralized data center. When duplicate data is discerned and discarded at the remote offices, only the minimal (unique) data from each branch has to traverse the WAN connections to the centralized backup server(s), as Figure 3 shows.

Figure 3. Enabling Diverse Protection Scenarios Through HP StoreOnce Federated Deduplication



Source: Enterprise Strategy Group, 2014.

To be fair, the discernment logic of client-side deduplication can incur a CPU and I/O increase on the production servers, so proper scaling is important. If done properly, the benefits to the overall data protection system are undeniably worth it.

Your backup software (agents and servers) as well as your storage have to share enough plumbing (APIs or other intelligent conduits/algorithms) to act in an integrated way.

Deduplication Optimizations and Markers

In addition to optimizing the algorithms and adjusting where in the backup flow deduplication discernment occurs, the most advanced (modern) deduplication technologies are beginning to offer other evolutionary capabilities as well.

Don't Rehydrate Across Devices or Sites

Modern deduplication solutions minimize the rehydration of data between devices. Consider an example of 500GB of data and even a minimal amount of deduplication (assumed 3:1 for discussion purposes only). After deduplication, the backup storage device may be storing less than 200GB, including multiple iterations of small changes.

- **In legacy storage-centric deduplication solutions**, making a replica of that data at another location involved rehydrating the entire data set and *transmitting all **500GB*** across the network to the target—even if it then ended up being deduplicated down to 200GB again.
- Often, **server-centric deduplication solutions** would only send the **200GB** (keeping the data deduplicated) because the data-movement logic in the backup server was deduplication aware. That's *much* better.
- But **client-side deduplication solutions** understand that on subsequent backups, if only 1GB of unique data was created in the production server, then only **1GB of new data** traverses through the backup server to the deduplicated storage, and only 1GB is replicated to the secondary site and its deduplicated storage.

One key to preventing rehydration (meaning that the data has to be un-deduplicated as part of moving it throughout the environment) is to ensure that the same deduplication logic is used throughout the data protection ecosystem. After it is deduplicated, it stays deduplicated. To be clear, this means that your backup software (agents and servers) as well as your storage have to share enough plumbing (APIs or other intelligent conduits/algorithms) to act in an integrated way.

Rapid Restores

Another key to successfully deploying dedupe is ensuring that restore operations are just as fast when performed from deduplicated backup storage as they are when performed from storage that does not leverage deduplication. Any deduplication solution that substantially increases the amount of time required to perform a recovery operation due to rehydration delays is going to adversely affect data recovery service level agreements (SLAs).

Fast (deduplicated) backup is nice. But it doesn't mean much without predictably fast (deduplicated) restores.

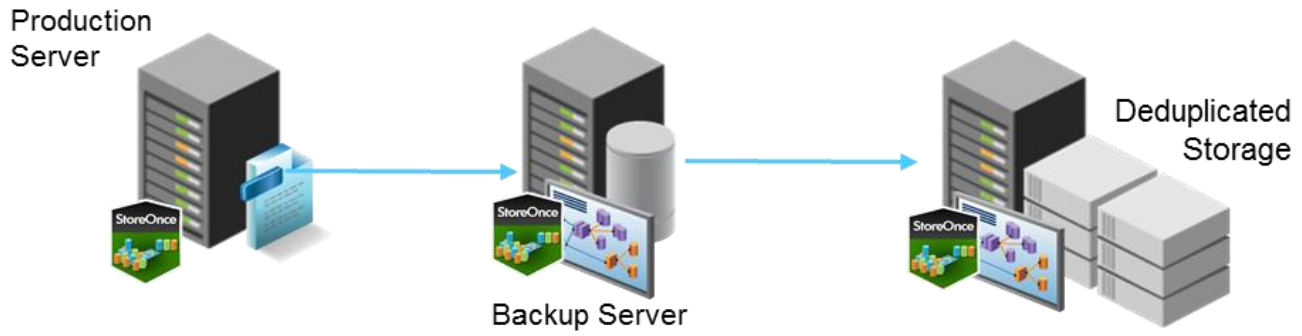
Fast (deduplicated) backup is nice. But it doesn't mean much without predictably fast (deduplicated) restores.

HP StoreOnce Backup

The Hewlett-Packard data protection portfolio, which includes both deduplication storage hardware (HP StoreOnce Backup), as well as backup software (HP Data Protector) and an ecosystem enabled by its APIs (HP StoreOnce Catalyst) is an example of "Dedupe 2.0," utilizing what HP refers to as ***Federated Deduplication***.

By using a consistent deduplication algorithm across the product suite (see Figure 4), HP StoreOnce minimizes problems related to different products with incompatible deduplication algorithms—problems that normally lead to comparatively slower recovery versus backup performance and decreased efficiency when scaling a deduplication solution to meet expanding capacity requirements.

Figure 4. HP StoreOnce Using Common Deduplication Technology from Source, to Backup Server, to Storage

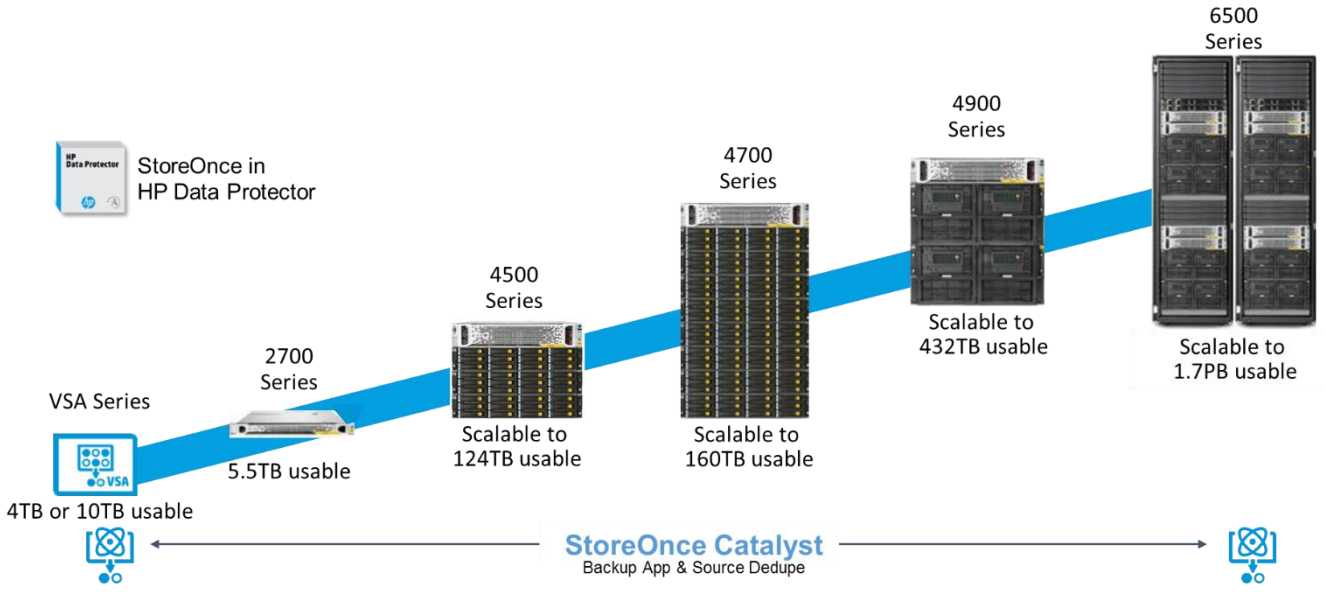


Source: HP, 2014.

HP StoreOnce Deduplicated Storage

The HP StoreOnce Storage Solutions are a family of storage arrays (see Figure 5). The largest component, the StoreOnce 6500 Backup system, is an enterprise deduplication appliance that uses a scale-out architecture. The 6500 can grow from a single 72TB usable two-node couplet to a capacity of 1.7PB usable. It also supports a pay-as-you-grow model that allows nodes to be added to the configuration without incurring downtime.

Figure 5. The HP StoreOnce Family



Source: HP, 2014.

Also notable is the highly resilient design of the HP StoreOnce 6500 to mitigate the impact of node failures. The product's Autonomic Restart feature configures backup jobs to restart after node failover without requiring the intervention of the data protection administrator. The StoreOnce 6500 is designed to automatically detect and remediate certain failures, hiding this complexity from data protection administrators and end-users.

Performance

According to HP, the HP StoreOnce 6500 Backup system is able to restore data faster than the rate at which it performs native backups. Table 1 shows HP's reported performance numbers for backups with the 6500.

Table 1. HP StoreOnce 6500 Performance Numbers, as Reported by HP

Action	Speed
Native VTL performance (max. configuration)	63TB / hour
With source-side deduplication via StoreOnce Catalyst (max. configuration)	139TB / hour
Native VTL performance (single couplet)	15.75TB / hour
With source-side deduplication via StoreOnce Catalyst (single couplet)	34.75TB / hour
Restore (VTL max. configuration)	75TB / hour

Source: HP, 2014.

According to HP, this performance places the 6500 in a strong position against its closest (unnamed) competitor:

- Native VTL performance (max. config.): **4.2** times its competition.
- With source-side deduplication (max. config.): **4.5** times its competition.
- Restore (VTL max. config): **5** times its competition.

The ability to perform deduplication at the source, at the backup server, at the appliance, and in HP Data Protector software gives organizations new options to help them minimize the growth of data under protection. In providing a solution that works at the source, backup server, and appliance level, HP offers all three options from Figure 2 (the deduplication flow chart), which is only possible because of HP's federated deduplication architecture.

Backing up the data is one thing. But as the amount of protected data grows, the ability to restore it fast enough to meet SLA agreements becomes more challenging. Notably, HP reports that its StoreOnce 6500 can ingest data (in VTL mode) *and* restore data at up to 75TB/hour. This is especially impressive because many deduplication technologies suffer an I/O penalty that causes them to restore much slower than their published ingest rate.

At the end of the day, however, the most important performance number is "price performance" ($/TB/hour). HP is claiming that it offers 60% better price performance than its primary competition.

HP StoreOnce VSA Virtual Storage Appliance

Along with continuing to grow its StoreOnce product line to support increasingly larger enterprise configurations, HP has released a virtual storage appliance (StoreOnce VSA) for StoreOnce. By providing a virtual appliance, something that ESG sees continually growing interest in,[2] several thought-provoking scenarios are enabled:

- Branch offices have the opportunity to deploy local backup/recovery solutions in locations that might have been judged to be "too small" to warrant a physical appliance. As part of consolidating and modernizing those branch offices with (perhaps) a single server-hypervisor host with all production assets running within virtual machines, the organization also can set up the backup server and storage appliance to be delivered virtually. In so doing, and because of HP's ability to replicate between appliances (with the data remaining deduplicated), the deduplicated data within the VSA can replicate from the branches to a larger physical HP StoreOnce platform at the data center.
- Cloud providers have an opportunity similar to the branch offices, where subscribers who don't already have a deduplication solution can utilize the VSAs for high-speed local recovery (with their existing backup software of choice), and then replicate the data via HP Federated Deduplication to the service provider's HP StoreOnce repository.

Conversely, cloud providers may choose to spawn individual HP StoreOnce VSAs per subscriber, so that each client has a completely autonomous deduplication appliance within the cloud—providing entirely separate security boundaries and a wider range of management options between the subscribers' IT staffs and the service provider's experts—without forcing subscribers to change their backup software (as a typical backup-as-a-service provider might do). Of course, there is a benefit to utilizing the HP software stack also, including HP Data Protector.

[2] Source: ESG Market Landscape Report, *Disk-based Backup Target Systems*, May 2013.

HP Data Protector Software

The HP backup, replication, and archiving story (referred to by many as HP BURA) would not be complete without considering the software capabilities of HP Data Protector. HP Data Protector (HP-DP) is designed to take advantage of the deduplication capabilities of the StoreOnce family and its Catalyst accelerator to enable client-side, media-centric, or target-based deduplication. In addition, by utilizing snapshots from HP storage arrays, HP-DP can perform "instant recoveries" of complex workloads or large datasets.

HP Data Protector also takes advantage of another HP acquisition: the Autonomy Cloud. Currently boasting 50PB, the HP Autonomy Cloud is a superset of what was once Iron Mountain Digital's online storage and recovery repository—and it is now an ideal location to replicate HP-DP data for offsite preservation. HP-DP also offers expanded hypervisor support (for VMware, Microsoft Hyper-V, and Citrix) as well as enterprise application protection for Microsoft, Oracle, and SAP server platforms—including granular data restore capabilities for Microsoft Exchange, SharePoint, and VMware.

Lastly, HP-DP also supports HP Autonomy's IDOL framework to deliver "meaning-based data protection," resulting in data protection, retention, and indexing based on the types and context of data in an environment.

HP StoreOnce Catalyst APIs

StoreOnce Catalyst is a software accelerator that enables deduplication anywhere, rather than just at those points in the network where specific vendor technologies allow it. StoreOnce Catalyst is the secret sauce in HP's *Federated Deduplication* architecture. It leverages a common deduplication algorithm across the enterprise and allows deduplication at the:

- Production source or "client."
- Backup or media server.
- Target appliance.

By enabling source-side deduplication, HP has the advantage of performing deduplication at the data creation point. Source-side deduplication eliminates the need for specialist deduplication hardware at remote and branch office sites. StoreOnce Catalyst allows customers to align backup with data protection needs, such as minimizing bandwidth utilization when moving data between sites or data centers. That said, it is important to note that source-side deduplication may not be ideal for every workload, nor for every IT topology. But by offering a single and consistent deduplication method, customers are free to choose source-, server- or storage-side deduplication intermixed throughout their environment, wherever each makes sense.

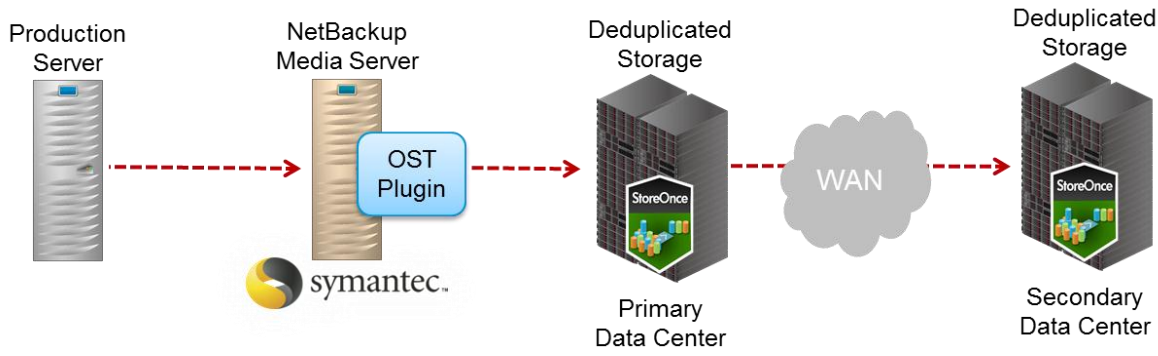
As a key "Dedupe 2.0" design goal, StoreOnce Catalyst provides a single technology that can be used in multiple locations on the network without requiring rehydration when data is transferred between source server, backup device, and target appliance (as depicted in Figure 2). As a software component, HP StoreOnce Catalyst supports HP Data Protector and an ecosystem of third-party backup software vendors through its HP StoreOnce Catalyst software development kit (SDK), including Symantec NetBackup via OST Integration.

HP is taking the journey even further with what it terms "*Federated Catalyst*," which provides even richer access across the Catalyst-enabled storage appliances. Capabilities include accessing the multi-node 6X00 series as a single pool—eliminating siloed capacity limits while maximizing performance with automatic load balancing and significantly reducing management overhead.

HP StoreOnce Partner Ecosystem—Including Symantec OST

The benefits of HP StoreOnce are not limited to Data Protector. One example is the integration of Catalyst with the Symantec OST plug-in. A NetBackup Media Server is able to utilize HP StoreOnce deduplicated storage, enabling smarter transit from the media server to the deduplicated storage and between HP StoreOnce arrays—without rehydration and managed wholly by NetBackup (see Figure 6).

Figure 6. HP StoreOnce Using Common Deduplication Technology from Source, to Backup Server, to Storage



Source: HP, 2014.

Although this design does not use StoreOnce Catalyst on the production server, the architecture still qualifies as a "Dedupe 2.0" solution. Symantec NetBackup has its own client-side deduplication and an "accelerator" feature for optimized discernment from the production node. It also utilizes the deduplication capabilities within the StoreOnce appliances.

The HP StoreOnce ecosystem includes BridgeHead Software and Oracle RMAN, and it is expected to grow even more as other backup software vendors embrace the Catalyst APIs for use with their products.

The Bigger Truth

Between the increased uses of server virtualization, the dynamic proliferation of private clouds, ever-growing unstructured data pools, and the advent of "big data," organizational data is going to continue to grow *a lot*. Effective deduplication reduces the impact of that growth, lessening the strain of constant storage growth on organizations. All organizations need an effective deduplication strategy. Ignoring deduplication will lead to data protection infrastructures being swamped with data.

Choosing a deduplication strategy involves assessing a range of factors including data type and location. By being able to support deduplication at the data source, backup server, and appliance, HP StoreOnce gives data protection solution architects a high degree of flexibility in developing a deduplication solution tailored to their needs.

By delivering a single solution through HP Data Protector paired with StoreOnce appliances, and a joint solution that offers StoreOnce capabilities through its ecosystem of Catalyst API-enabled partners, HP has taken aim at a unified and broadly applicable data protection solution that challenges the presumptions of deduplicated storage. With the addition of cloud capabilities and the "understanding" of data through Autonomy's assets, the data protection game just got more interesting.

Deduplication is not new—and HP was certainly not the first to market it. Instead, by watching how deduplication was introduced and listening to the evolving demands of customers who struggle with storage and backup issues, HP built StoreOnce as a next-generation or "Dedupe 2.0" architecture that is available now. With its formidable enterprise experience, server and storage product lines, and broad partner ecosystem, HP intends to catch up with and, in fact, surpass the status quo to bring better deduplication at a lower cost.

If you haven't already, now is the time to adopt deduplication fully within your data protection strategy. And you may want to reconsider a name that many view as synonymous with enterprise servers and storage: HP.

Enterprise Strategy Group | **Getting to the bigger truth.**

20 Asylum Street | Milford, MA 01757 | Tel: 508.482.0188 Fax: 508.482.0218 | www.esg-global.com

4AA4-1782ENW