White Paper

HP StoreOnce Is "Better Together" with HP Data Protector 7

By Jason Buffington, Senior Analyst

December 2012

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

Contents

HP Is Upping the Ante on Deduplication

In June 2012, HP released StoreOnce Catalyst enhancements to its new top-end B6200 appliance (released by HP in November 2011) with support for HP Data Protector 7 backup software. Also released was the StoreOnce Catalyst API software toolkit for enabling third-party backup software to better leverage the capabilities of HP's StoreOnce architecture.

In December 2012, HP upped the ante again by providing a firmware update for the full range of existing StoreOnce appliances to further improve their performance. The company also added StoreOnce Catalyst deduplication capabilities across the entire StoreOnce family portfolio.

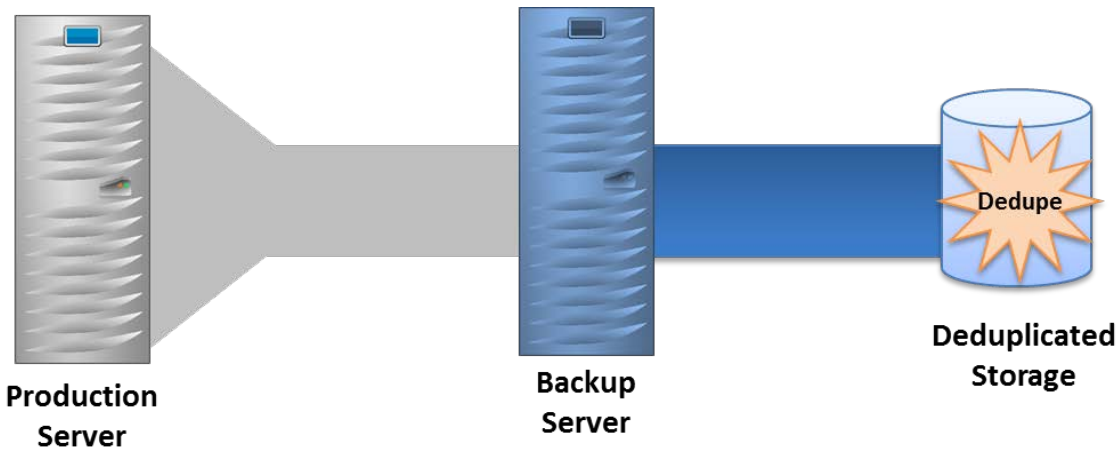
During the initial launch, ESG published a white paper detailing its perspectives on where data deduplication, as part of disk-based backup, was headed in the industry and the applicability of the HP StoreOnce solution in its approach with "Dedupe 2.0." That white paper examined the HP StoreOnce portfolio and its approach to deduplication; this paper will look closer at the benefits to be realized in the combined scenario of HP StoreOnce appliances being used with HP's own Data Protector 7 backup software.

Perhaps the most visible evolution in deduplication is related to the subject of "where" the deduplication occurs. Instead of simply taking place within a target storage device for economic efficiency, deduplication often now occurs within the backup server or even on the production server(s) themselves, as seen in Figures 1, 2, and 3.

Storage-centric Deduplication

Figure 1 shows **storage-centric deduplication**—whereby any deduplication-capable storage device (on right) is connected to a backup software solution that is unaware of the storage solution's capabilities—requiring the backup software to have no co-engineering or integration work. In a storage-centric deduplication configuration, all data (including redundant copies and chunks) is sent from the production server to the backup server, then on to the storage solution. Unbeknownst to the source or backup server, redundant data elements are discarded within the storage solution for more effective storage of disk-based backups.

Figure 1. Storage-centric Deduplication—to Optimize Storage Economics



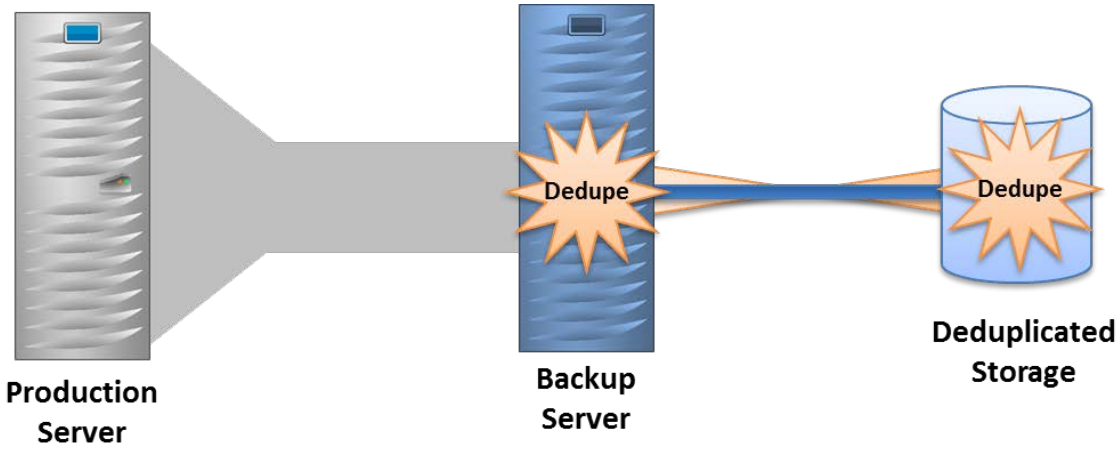
Source: Enterprise Strategy Group, 2012.

In what HP refers to as "target-based deduplication," the merit of storage-centric deduplication is the compatibility that it offers with the widest variety of backup, archiving, and other data-movement technologies because those products access the storage device via their normal means—without awareness of the optimizations within the array.

Backup-Server Deduplication

Figure 2 shows **backup-server deduplication**—whereby the backup server software is aware of what is in the deduplicated storage solution, so that redundant data (or partial chunks) are not sent from the backup server only to then be discarded by the storage. Instead, the backup server is aware of what data is within the deduplicated storage pool, thereby improving its discernment of what should be stored or discarded.

Figure 2. Backup Server-centric Deduplication—to Optimize the Backup Server's Network and Storage



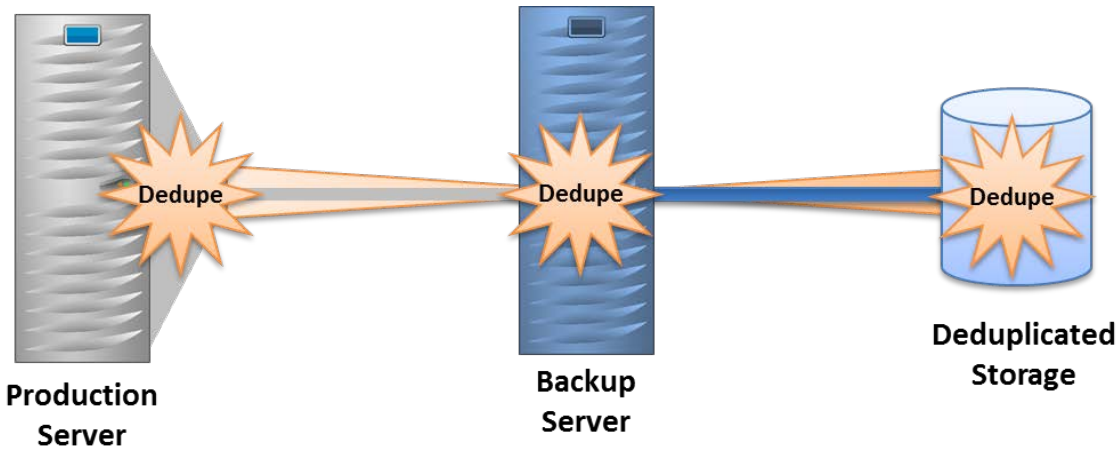
Source: Enterprise Strategy Group, 2012.

By enabling deduplication at the backup server, customers can continue using their previously installed backup software but gain better performance and economics by using deduplicated storage in a smarter way (and not blindly). It is worth noting that although this enablement improves many environments, all of the production data being backed up, including the redundant elements, is still sent from the production source to the backup server, even though the network and processing aspects between the backup server and the storage are optimized.

Source-side Deduplication

Figure 3 shows **source-side deduplication**—whereby the production server discerns not to send redundant elements that would be discarded. This approach maximizes the infrastructure benefits of deduplication because no additional network bandwidth or processing is required through the backup server.

Figure 3. Source-side Deduplication—to Optimize Overall Deduplication Across the Environment



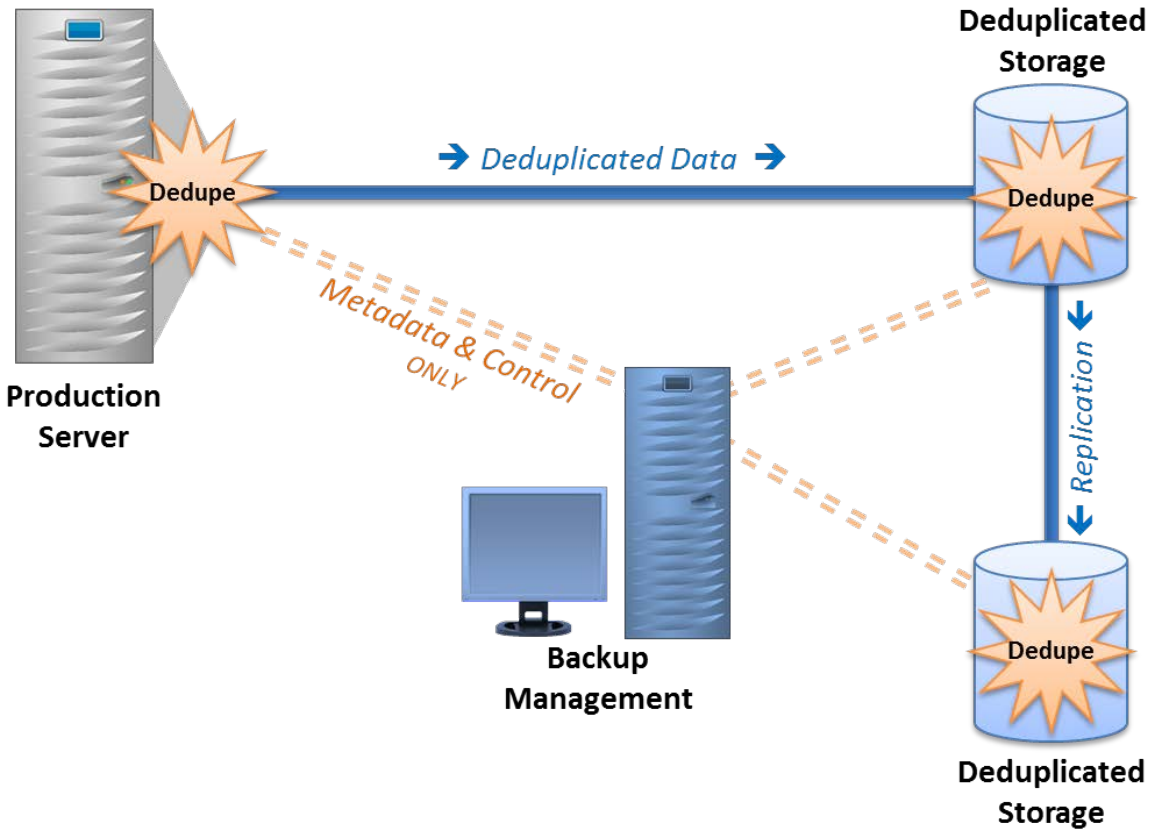
Source: Enterprise Strategy Group, 2012.

Assuming viability by the production workload, as well as interoperability by the backup software and hardware, source-side deduplication is often perceived as the ideal model for many environments and a logical next step in

the evolution of data protection overall. The answer of "where" deduplication should occur (source-side, backup-server, or storage-centric) isn't always a simple or unilaterally correct answer.

In fact, ESG believes that as deduplication architectures continue to mature, those users choosing to leverage source-side deduplicated data streams may in fact circumvent a backup server entirely—with the data being transmitted directly to the backup storage appliance, and the backup server simply storing the metadata for indexing by performing job scheduling and monitoring, as well as initiating restore operations (see Figure 4).

Figure 4. Optimized Data Movement with Backup Management, Without an Intermediary Backup Server



Source: Enterprise Strategy Group, 2012.

Choosing Where Deduplication Should Occur

This is not to say that source-side deduplication is best for all scenarios or customers. Reasons certainly exist regarding why it is more optimal for deduplication to occur on the backup server, such as the ability to offload any incremental I/O or CPU processing from a mission-critical and highly utilized production platform, or compatibility with existing backup software that is thoroughly deployed across large environments.

Many environments will often consider a hybrid approach, where source-side deduplication is used when possible, and backup-server deduplication is used for other scenarios where source-side isn't viable—with a preference toward utilizing software and hardware deduplication in complement to each other. But in all cases, deduplication is considered a requirement to effectively utilize disk-based backup solutions from any vendor.

How to Deduplicate

As important as the question about *where* deduplication happens (source-, backup-server-, or storage-centric) is *how* the storage is accessed by the backup application—*how* deduplication occurs, and *how* the data behaves once deduplicated.

How the Storage Is Accessed by the Backup Application

For the broadest compatibility of backup applications, deduplication storage solutions will present themselves as either "generic" disk or tape devices:

- Almost every backup software platform can write to a simple **network-attached storage (NAS)** device via either CIFS or NFS file shares, which then performs its deduplication operations within the storage solution.
- Similarly, by the storage presenting itself as a **virtual tape library (VTL)**, backup solutions that utilize tape hardware can store data within the deduplicated solution with the belief that the data is stored in a library of tape cartridges.

Although these approaches to accessing the deduplication appliance provide generic disk- or tape-based access to the deduplicated storage, they typically require the deduplication to be performed strictly within the storage appliance because the backup software believes that it is simply writing to "just" a NAS file share or tape device.

For more optimal performance and manageability, the backup server and/or even the backup agents on the production servers need to be dedupe-aware:

- Storage solutions that support **APIs that are usable by the backup server** (such as HP StoreOnce Catalyst) enable one's backup servers to be aware of what is already in the deduplicated storage pool, therefore sending less data between the media server(s) and the storage solution and enabling a backup-server capable deduplication solution.
- Storage solutions that provide **APIs that are usable within the backup agents on the production server** can achieve even greater efficiencies by discerning what should be deduplicated at the production server (or the backup server) instead of within the storage appliance.

HP StoreOnce Catalyst

Any backup software could leverage an HP StoreOnce appliance as a NAS or VTL and gain storage-centric deduplication benefits, but the optimized experience is enabled through the HP StoreOnce Catalyst APIs that improve performance and manageability:

- HP has published the Catalyst APIs so that *third-party backup solutions* can choose to enable their backup servers to discern their own deduplication through software integration with the StoreOnce appliances.
- *HP's Data Protector 7* takes this "better together" scenario even further, enabling source-side or backup-server enabled deduplication through its full utilization of the StoreOnce Catalyst APIs.

StoreOnce Catalyst essentially adds intelligent communication between the backup software and the storage device—enabling either backup-server or source-side deduplication and enabling the backup software application to manage the core functions of the appliance in a cohesive way.

How the Deduplication Occurs

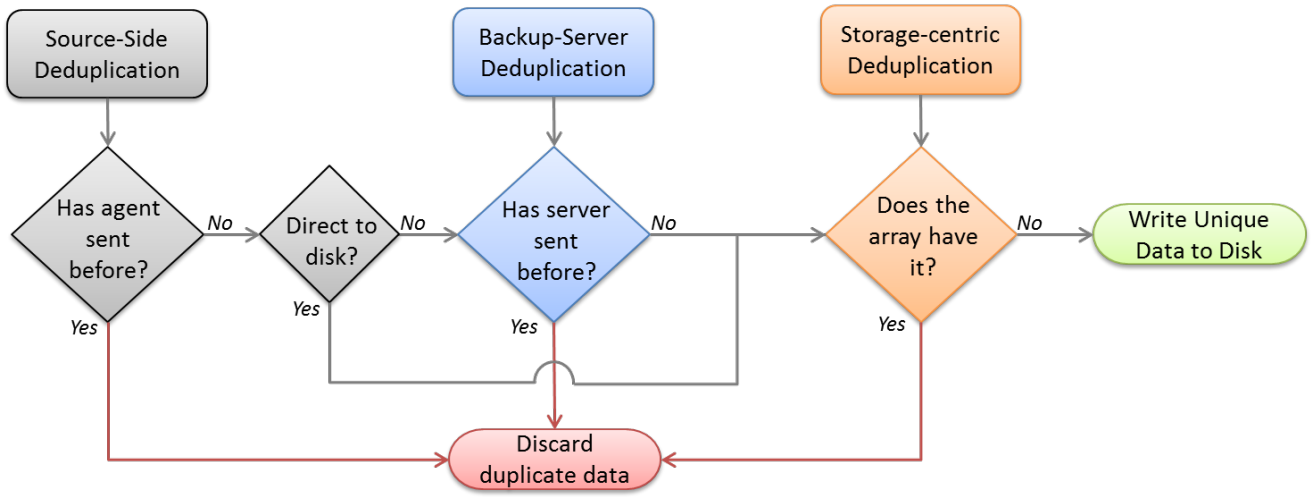
The most common method of deduplication is to deliver *inline processing*: As data is streamed to the storage solution, the discernment whether to retain or discard the data chunks occurs in near real-time. The alternative to inline processing requires all data to be written to the array. Comparisons and deduplications then occur through a method called *post-process*. HP StoreOnce appliances utilize inline processing by leveraging high computational processors within the array's controller nodes, so that the identification of "deduplicate-able" data, as well as any metadata indexing, compression, or encryption, can occur without appreciably slowing down the flow of data into

or out of the storage solution. Essentially, as the deduplication appliance receives the blocks of data to be written, it takes each chunk or block of data ranging from 4K to 32K and compares it against the chunks that already reside in the storage appliance.

- If the chunk *already exists* within the storage, the duplicate chunk is discarded without writing anything—and a pointer or metadata entry notes where the duplicate block would have been.
- If the chunk *is not yet in the storage appliance*, it is written to disk, and its metadata is added to the catalog in order to help identify duplicates in the future.

Figure 5 shows a simplified view of how data is discerned as duplicate (or not).

Figure 5. Where Deduplication Occurs, and Its Data Movement



Source: Enterprise Strategy Group, 2012.

How Data Behaves After Deduplication

Imagine a model where data is deduplicated and stored, but when it is moved to another device, it must first be "rehydrated." That process causes slower performance during the movement and affects the solution's ability to process other backup or restore requests. Instead, contemporary deduplication solutions ensure that once data has been deduplicated, it stays in its optimized form as it moves throughout the infrastructure for purposes such as multi-site replication for disaster recovery or multi-access for development/testing.

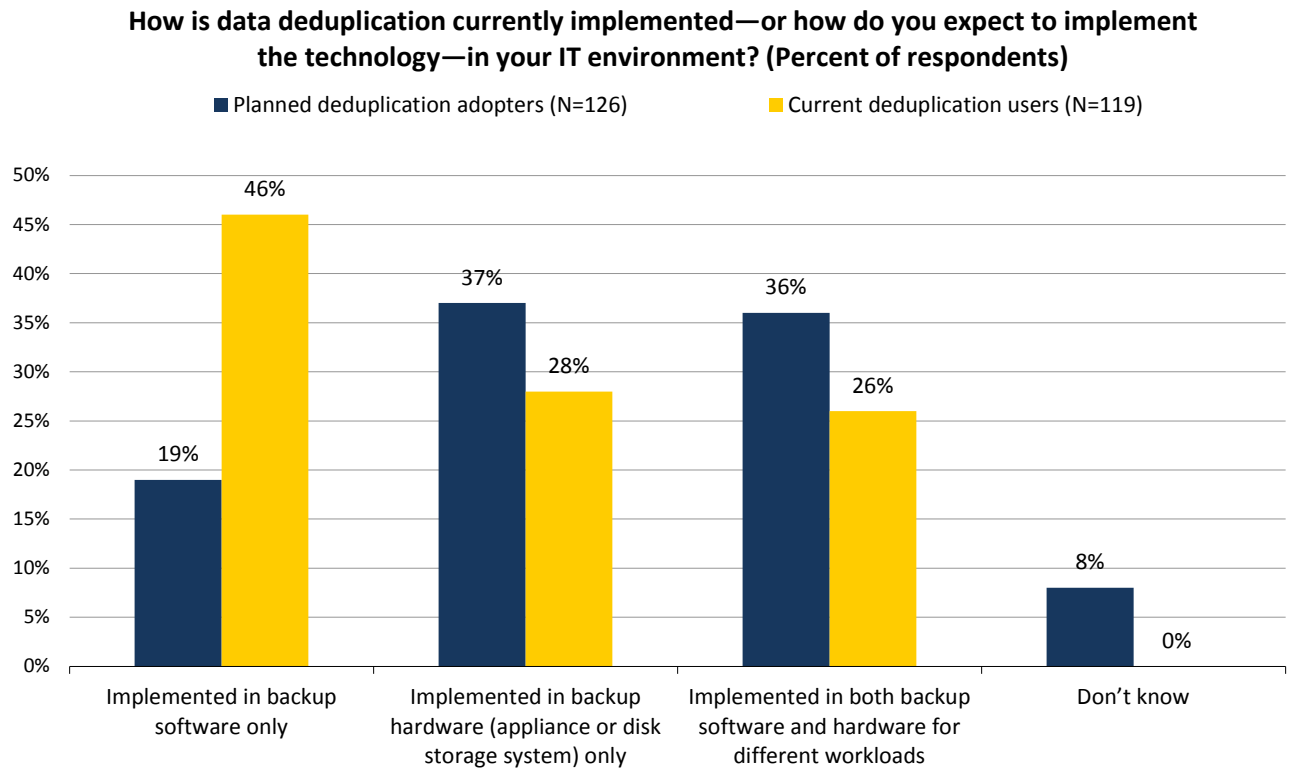
As seen earlier in Figure 4, this method often changes the overall architecture of one's backup solution by encouraging data movement between production sources and storage repositories (plural), even without the need to go through an intermediate "backup server"—all while the data remains in its optimized form.

How to Achieve Deduplication in Software, Hardware, or Both

ESG research shows that the choice of data deduplication is not a matter of *if* but *when*—as an inevitable part of every disk-based data protection solution. In fact, the only questions left to answer on deduplication are how and where that optimization should occur. In a recent research study (see Figure 6), ESG found an appreciable disparity between where current users were performing deduplication versus where those who were planning to use deduplication expect to implement it.[1]

[1] Source: ESG Research Report, *Trends in Data Protection Modernization*, August 2012.

Figure 6. Current and Planned Deduplication Methods



Source: Enterprise Strategy Group, 2012.

Of the ESG respondents who were *currently using deduplication (shown in yellow):*

- Nearly half (46%) reported using strictly software-based deduplication.
- Twenty-eight percent reported using exclusively hardware-based deduplication.
- Twenty-six percent reported using a combination of hardware and software.

With many customers presuming that existing hardware approaches were cost-prohibitive, it is not surprising to see nearly half of all deduplication occurring strictly through software implementation today. But as new deduplication hardware alternatives (such as the HP StoreOnce family) as well as new optimization scenarios (such as software acceleration to hardware-based deduplication) continue to be introduced, it is reasonable that many would reconsider solutions that are not software centric.

For the ESG respondents who were *not yet using deduplication but planning to (shown in blue),* the preference toward using hardware-based deduplication is dramatically higher:

- Thirty-seven percent plan to implement deduplication exclusively via hardware.
- Thirty-six percent plan to use software+hardware-based deduplication.
- Only 19% plan to use a software-only approach.

For HP customers, "*exclusively hardware*" would imply deduplicating within a StoreOnce appliance. Meanwhile, "*software+hardware*" would imply using StoreOnce catalyst within either the backup-server software or the source-side agents to optimize the deduplication capabilities and the manageability of the overall solution.

Putting the HP Pieces Together

To enable software-based source-side deduplication (what HP refers to as "Application Source Deduplication"), HP's StoreOnce Catalyst API development kit empowers backup software to fully utilize its StoreOnce hardware. As proof of the power of Catalyst's capabilities, HP also announced that version 7 of HP Data Protector is StoreOnce Catalyst-enabled (with other vendors announcing support of the Catalyst APIs within their future products).

By embedding StoreOnce Catalyst within its DP7 agent, HP is able to consistently utilize a single deduplication methodology across the entire backup infrastructure, including:

- The backup agents on the production servers for source-side deduplication.
- The DP7 backup server for indexing, cataloguing, and job management for backup server deduplication.
- The StoreOnce backup (storage) appliance for deduplication within the storage solution.

The integration of Catalyst across DP7 and the StoreOnce appliances enables some interesting management capabilities as well:

- Control of where the deduplication occurs (source, server, or storage) from within the DP7 console.
- Centralized management of any additional StoreOnce appliances, as well as long-distance replication for protection or disaster recovery.
- Variable retention policies for the data at the backup server, the primary StoreOnce appliance, and any secondary StoreOnce appliances for multi-site flexibility and recovery scenarios.

According to HP, alternative deduplication solutions are built from components and products that may interoperate but are unable to utilize a consistent deduplication methodology. Because of this circumstance, some solutions may efficiently store data within their part of the infrastructure but force rehydration of the data as it moves from one "island" to the next, between backup server(s) and storage array(s).

And while not included within the StoreOnce deduplication scenario yet, it is important to note the single-vendor advantage of data protection management that comes from the scenario above (source agents to backup server to storage appliance) also includes HP tape solutions for long-term retention.

StoreOnce Enhancements for 2013

In December 2012, HP announced StoreOnce Catalyst firmware enhancements to the StoreOnce single node storage appliances that are available to existing HP customers. The update changes the core behavior of earlier HP StoreOnce D2Dxxxx series appliances to support StoreOnce Catalyst, thus enabling those smaller StoreOnce D2D appliances to interoperate with the StoreOnce B6200 large appliance, as well as with HP Data Protector 7—bringing HP StoreOnce capabilities to small and midsize organizations. Although a destructive wipe over the existing data is required, the benefits are worth it. After updating, HP boasts a native performance increase of up to 2X, with Catalyst performance improving up to 3X—essentially offering HP customers even more performance for the same price.

"Right-Sized" Distributed Backups with Unified Deduplication

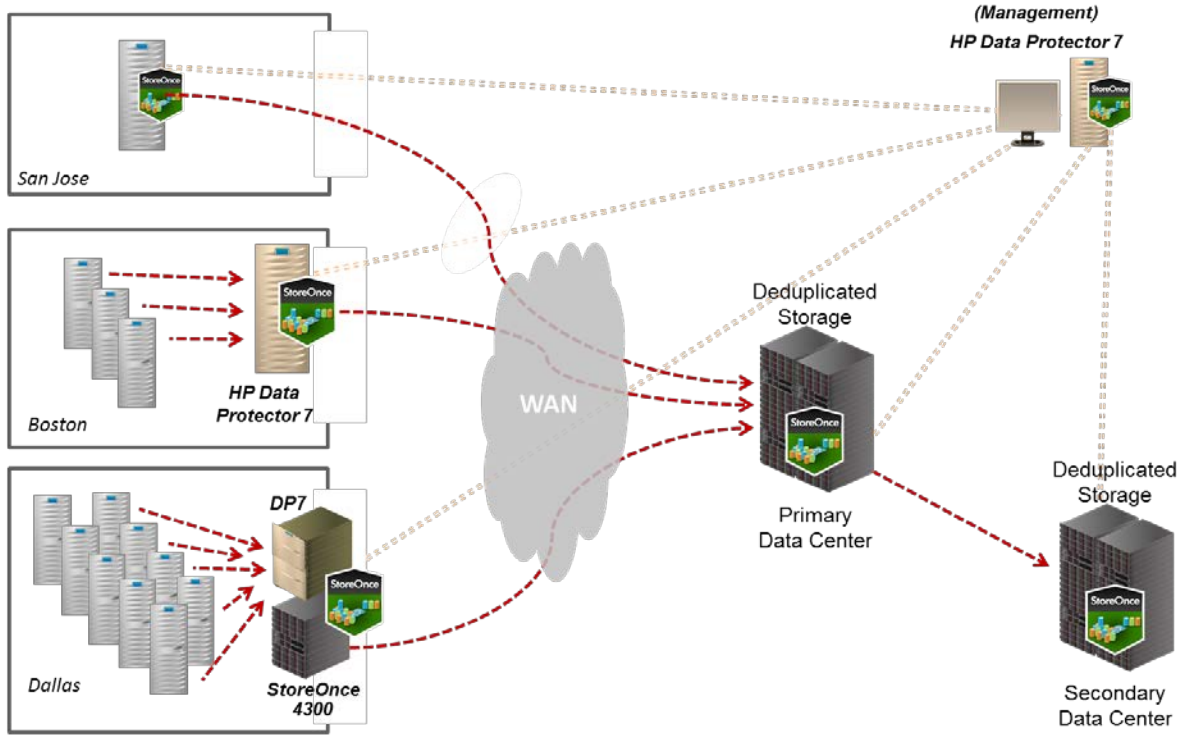
The plight of remote or branch offices (ROBOs) and other distributed environments is a key scenario benefitting from this homogenous approach to deduplicated data protection. It has long been a goal of IT to centralize backup of remote office data so that remote employees would neither be managing tapes nor monitoring backup jobs. It is more reasonable that an organization's central IT team would manage the backup process—with as limited interaction with remote personnel as possible, and as limited remote server impact as possible.

Today, there are caching solutions, replication technologies, backup software optimizations, virtualized backup appliances, and a myriad of other attempts to solve centralized backup goals. However, the answer that most IT experts continue to desire and look for centers on efficiently utilizing the same tools for their remote data that they

trust in the data center. The reason? Every other approach adds new technologies and management requirements to an already-taxed IT team.

The nirvana (presumed perfect) scenario has always been to utilize data center data protection methods without taxing the intra-company WAN. This is a scenario where a unified deduplication approach truly shines. By leveraging source-side deduplication, HP DP agents are able to discern what has already been protected and stored within an HP StoreOnce appliance—not only from that particular remote data source, but also from any other data source whose data is protected by the same StoreOnce appliance (see Figure 7).

Figure 7. Complete HP StoreOnce and DP7 Solution Across Branch Offices and Data Centers



Source: Enterprise Strategy Group, 2012.

Figure 7 shows three remote offices:

- A small, single-server office can protect its data directly to the primary data center's deduplicated storage.
- A midsize remote office first protects its data to a local HP Data Protector backup server (before replicating the data to the StoreOnce appliance within the corporate data center).
- Lastly, a large office can now have a smaller StoreOnce storage appliance and DP7, before replicating its data to a larger appliance at the data center.

Recognizing that this distributed architecture consolidates so much data within the data center's appliance, HP's largest B6200 StoreOnce offerings operate as clustered pairs to ensure availability of the backup repository. The distributed replication also enables the data center to provide data survivability or disaster recovery capabilities for the remote offices, while the primary data center's data is then replicated to a secondary data center. And, regardless of how the data gets to the data center, the data remains in its deduplicated state throughout.

The alternative to StoreOnce's universal deduplication approach might require multiple optimization mechanisms:

- First between the backup agents and the backup servers.
- Then, different deduplication or compression activities between the backup server and the storage (or from onsite storage to offsite storage).

Other alternate approaches also might require a backup server (physical or virtual) to operate at each remote site so that optimized backup-server-to-backup-server communication would traverse the WAN—but that would consume additional CPU, I/O, and storage at each remote site.

By using the same deduplication methodology across the infrastructure, HP's "unified protection" solution ensures that only the unique data fragments that are discerned from the HP DP 7 agents on the production servers will travel across the low-bandwidth links between sites. Only those unique fragments will be indexed through the Data Protector server and stored within the StoreOnce device. And only those unique fragments will be replicated from one data center's StoreOnce appliance to an offsite StoreOnce appliance for disaster recovery. In all cases, because the deduplication methodology is shared among the StoreOnce hardware array(s), StoreOnce Catalyst, and the DP7 agents, the backed up data never has to be rehydrated or translated between components of the backup infrastructure.

Other Considerations for an All-HP Solution

Aside from the "unified protection" capability that is exemplified by the remote office solution described above, other notable benefits of an HP StoreOnce plus Data Protector 7 solution exist.

Finding New Meaning in Backup

One of the most interesting aspects of a DP7 plus StoreOnce solution is a capability that HP refers to as Intelligent Data Operating Layer (IDOL), which provides contextual search based on data recognition. Using this technology, HP customers could potentially save thousands, if not millions, of dollars by leveraging their backup data repository for e-discovery or other litigation processes through the DP7 backup software.

Is It Faster?

Although having a common methodology for deduplication has some great theoretical appeal, and logically, it certainly stands to improve scenarios such as centralized backup of remote offices—it won't matter if the deduplication storage performance isn't viable.

According to HP, a fully configured HP StoreOnce B6200 cluster can achieve backup speeds of up to 100TB/hour (with Catalyst) and restore speeds of up to 40TB/hour—numbers that greatly exceed the published statistics of most competitive technologies. Historically, ESG has found HP to be highly conservative in its marketing, particularly in regard to performance metrics. We have not yet independently confirmed the B6200's performance measurements, although we hope to do so in the near future. In the meantime, recognizing the brashness of HP's reported performance and the lack of an industry-standard method for measuring deduplication efficiency, ESG suggests that enterprises that need superior backup performance (and any organization that is looking for deduplication alternatives) test the StoreOnce solution, whether they are using HP Data Protector or using other backup solutions that can leverage the HP StoreOnce hardware.

Is It Better?

Without quantifiable comparisons of performance, two other measures do bear consideration: manageability and price-performance.

By entering the deduplicated storage market after its initial inception, HP has been able to listen to feedback from customers of competitive solutions regarding ease of deployment and ease of use. With those ease-related goals in mind, HP claims 50% fewer clicks for initial configuration and a "single pane of glass" experience through DP7 for

the ongoing experience. ESG believes that from an ongoing management standpoint, most deduplication storage solutions should be relatively "fire and forget," other than when troubleshooting is needed. So the ability to monitor and manage the health and effectiveness of the StoreOnce hardware through the DP7 backup software is meritorious and "somewhat unique" (i.e., it's on a short list). That type of unified management capability often requires a single-vendor solution for both the backup software and the hardware.

By assuming that ingest rates (as well as deduplication efficiencies) may vary per environment, ESG instead suggests a price-performance metric that assumes a relatively equitable deduplication efficiency across competitors yet focuses on cost per capacity of usable (non-deduplicated) storage. HP DP7 is typically not presumed to be a low-cost software solution. But its combination with StoreOnce hardware—which costs 30% less than comparable deduplication storage systems, according to HP—may make the overall data protection infrastructure markedly less expensive than some other alternatives.

The Bigger Truth

Growth of unstructured (file) data, sprawl of virtual machines' disks (as virtualization becomes a commodity), and the ever-growing demands and scale of applications (including the advent of big data) are all source-side data protection challenges.

Certainly, deduplicated arrays make it cheaper to store backup data. By adding deduplication awareness to one's backup server(s), one can store data more cheaply *and* operate backups more efficiently. That efficiency grows considerably when optimizations and deduplication discernments occur not just on backup server(s), but also on production source servers.

HP started redefining the alternatives to deduplicated backups through its introduction of HP StoreOnce backup appliances, and it has continued to evolve its offerings through a Catalyst-enabled partner ecosystem and now its introduction of Data Protector 7 for an all-HP solution. In doing so, HP is promising a "single pane of glass" manageability experience with reported backup- and restore-performance numbers (as well as cost effectiveness levels) that currently are unmatched in the market.

ESG believes that as much as deduplication originally solved the main setback for disk-based backups (the cost of storing the data), the evolution of deduplication must bring the optimization discernment closer to the source data (and not just within the storage itself). With the "Dedupe 2.0" strategy that HP has built into its StoreOnce product line, coupled with partners that utilize Catalyst as well as HP DP7, there appears to be a new contender for "optimized" deduplicated backups that should be considered, evaluated, and certainly watched for future innovation.

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