

White Paper

HP StoreOnce Deduplication Software

Technology Fueling the Next Phase of Storage Optimization

By Lauren Whitehouse

June, 2010

This ESG White Paper was commissioned by Hewlett-Packard and is distributed under license from ESG.

Contents

Introduction	3
Overview	3
Data Deduplication	4
HP’s Deduplication Vision and Solutions	5
HP StoreOnce.....	6
HP D2D Backup Systems.....	6
Deduplication Purchase Criteria	7
HP Converged Infrastructure and HP StoreOnce	9
The Bigger Truth	10

All trademark names are property of their respective companies. Information contained in this publication has been obtained by sources The Enterprise Strategy Group (ESG) considers to be reliable but is not warranted by ESG. This publication may contain opinions of ESG, which are subject to change from time to time. This publication is copyrighted by The Enterprise Strategy Group, Inc. Any reproduction or redistribution of this publication, in whole or in part, whether in hard-copy format, electronically, or otherwise to persons not authorized to receive it, without the express consent of the Enterprise Strategy Group, Inc., is in violation of U.S. copyright law and will be subject to an action for civil damages and, if applicable, criminal prosecution. Should you have any questions, please contact ESG Client Relations at (508) 482-0188.

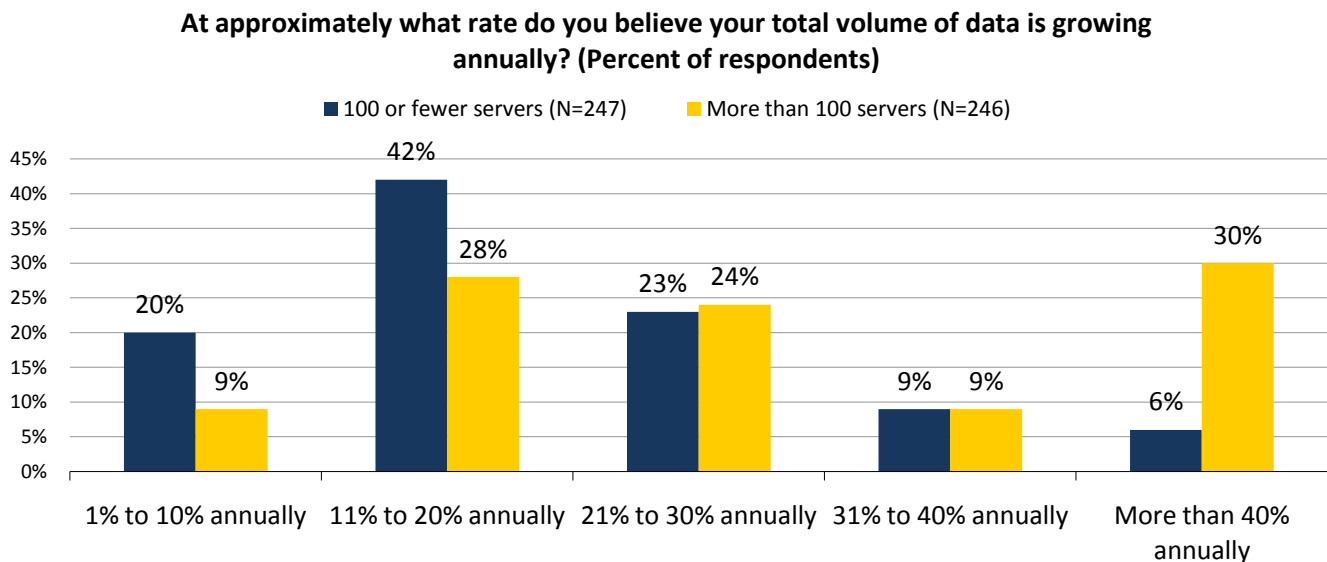
Introduction

IT organizations are plagued by issues of data growth, more stringent requirements for recovery time objectives (RTOs) and recovery point objectives (RPOs), and a reduction in operational staff to manage it all. Disk-assisted data protection aids in accelerating backup and recovery performance—and data deduplication makes the economics of implementing disk-to-disk backup more feasible. Today’s most popular backup storage targets featuring deduplication, however, don’t address the long-term operational issues that are sure to arise as secondary storage environments expand; namely, scale, performance, robustness, flexibility, and ease of management. HP’s StorageWorks division offers deduplication in its D2D Backup System family today with a vision and a delivery strategy to meet the demands of IT organizations in the future. This paper examines trends and customer requirements in storage regarding data deduplication and details how HP is best positioned to deliver highly efficient deduplication solutions based on its StoreOnce deduplication technology.

Overview

Several forces are converging to fuel interest in and adoption of deduplication technology for data protection: unprecedented data growth, inefficiency in data protection applications, and increased use of disk for secondary storage. ESG research found that most (58%) organizations are experiencing annual growth rates of 11% to 30%. However, as shown in Figure 1, organizations with more than 100 servers are experiencing growth rates beyond that mark, with 30% of those with more than 100 servers experiencing annual growth of 40% or more.¹

Figure 1. Data Growth Rates, by Quantity of Production Servers



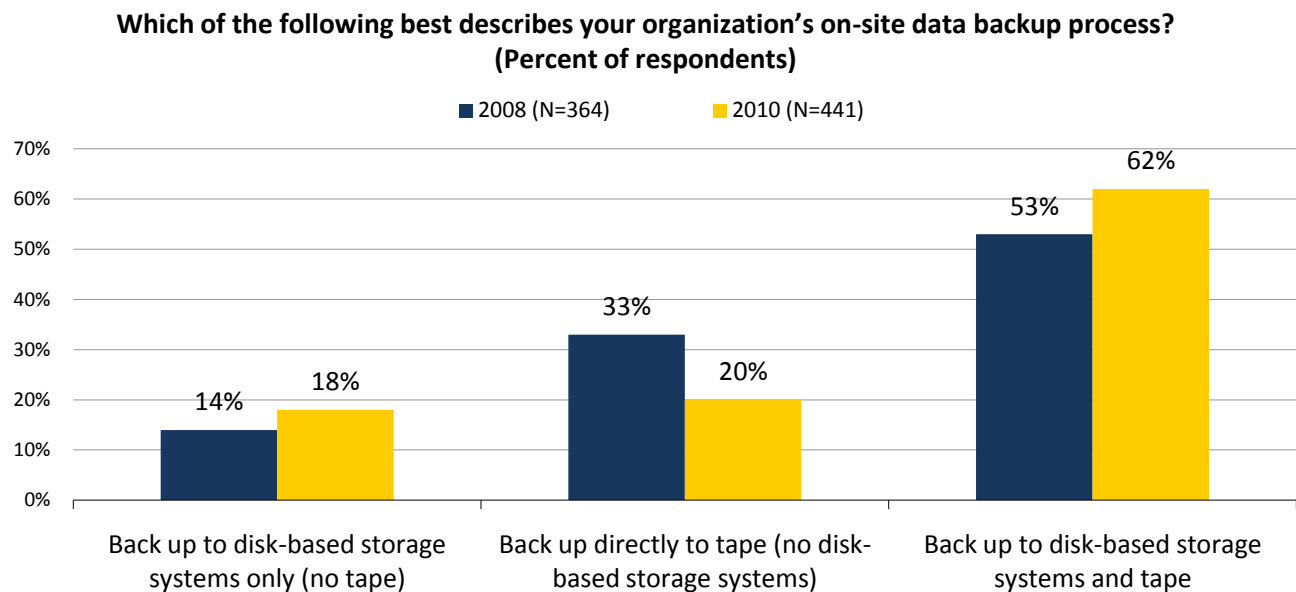
Source: Enterprise Strategy Group, 2010.

Data redundancy in secondary storage is a result of inefficiency in data protection applications. Most backup solutions make multiple copies of the same file despite the fact that only a small amount of the data within the file may have been changed. Multiple copies of data are made for recovery processes (at least once per day, but sometimes multiple times daily). Copies of the copies are sent to an offsite location for disaster recovery (DR) purposes. All told, dozens of copies of the same data could be stored for extended periods of time, depending on retention policies.

Over the last several years, disk has been increasingly used as the initial or final resting place of backup copies, improving the performance and reliability of backup and recovery operations. Today, 80% of ESG research respondents cite the use of disk in backup processes. As shown in Figure 2, over the last two years, disk-to-disk backup increased 15% and disk-to-disk-to-tape backup increased 22% while disk-to-tape backup decreased 65%.

¹ Source: ESG Research Report, [Data Protection Market Trends](#), April 2010. All statistics come from this report unless otherwise specified.

Figure 2. Use of Tape and Disk in Backup Processes – 2008 vs. 2010



Source: Enterprise Strategy Group, 2010.

While disk-based data protection can contribute to lower operational expenses and improvements in backup and recovery service level agreements (SLAs), it can also increase capital expenses. Organizations are, however, motivated to gain control over the situation. The flood of data has resulted in increased storage system costs; growing difficulty in providing adequate levels of data protection; and mounting stress on data center power, cooling, and floor space. Many IT organizations are turning to capacity reduction technologies, such as data deduplication, to solve these problems. In fact, 38% of ESG research respondents are currently using deduplication and another 40% plan to in the next 12-24 months.

Data Deduplication

Deduplication identifies and eliminates redundancy by writing only unique data to storage. When duplicate files, blocks, or byte sequences are detected, only a pointer linked to the unique piece of data is stored. This pointer consumes significantly less capacity than storing the whole item multiple times.

When used in backup and recovery, deduplication changes the economics of disk-based data protection. Eliminating redundancy minimizes storage capacity requirements, which can stem or slow the purchase of disk capacity. For example, a 10:1 reduction ratio means that 10 times more data is protected than the physical space required to store it and a 20:1 ratio means that 20 times more data can be protected. Factoring in data growth, retention, and assuming deduplication ratios in the 20:1 range, 2 TB of storage capacity could protect up to 40 TB of retained backup data.

The economic savings don't end there. Deduplication also optimizes network bandwidth, enabling replication with lower bandwidth requirements. Reducing the volume of data transferred over a WAN connection makes backup consolidation from remote locations more feasible. Similarly, electronically transferring data off site for disaster recovery purposes makes automated electronic vaulting less time-consuming and costly.

The impact of deduplication doesn't end with capacity and network optimization. Deduplication facilitates higher levels of service and it can increase the likelihood that more data can be rapidly backed up to and recovered from disk, meeting prescribed backup windows and providing improved recovery time objectives (RTOs). Shortening backup time may result in enabling more frequent backup copies on disk, leading to improved recovery point objectives (RPOs). It also allows organizations to increase retention policies for data on disk. Today, 68% of research respondents claimed retention of data on disk in disk-to-disk-to-tape strategies for *one month or more*, while 66% of respondents in 2008 retained data on disk for *only one week*.

Organizations are also likely to see operational improvements; deduplication contributes to time savings by enabling more data recovery to occur from disk versus an operator-intensive and time-consuming tape process. It also reduces the footprint of disk—reducing power and cooling costs as well as minimizing data center floor space consumption.

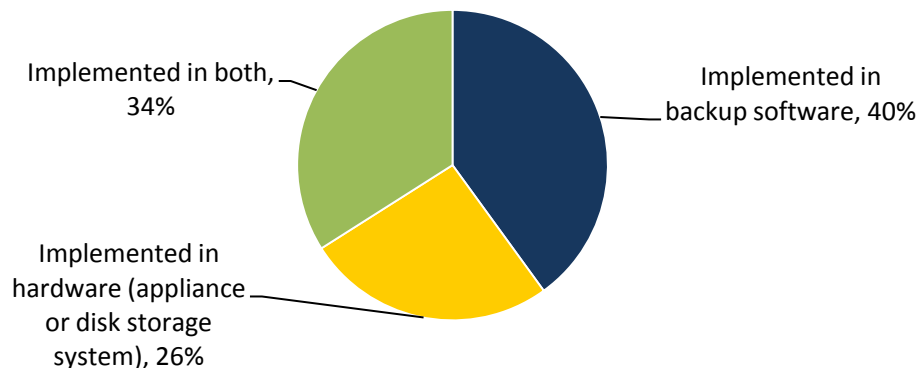
HP's Deduplication Vision and Solutions

HP has been delivering data deduplication technology for a few years via its disk-based backup portfolio. With its StoreOnce deduplication engine,² HP is undertaking a bold, multi-phase rollout of a new generation of deduplication technologies to address what it sees as the shortcomings of “point solution” deduplication offerings available today: high costs, complexity, high management overhead, rigid solution stacks, scalability limitations, and silos of heterogeneous deduplication storage.

HP's claims of high fragmentation and point solutions are not unfounded. In addition to backup solutions, data deduplication is implemented in primary and archive storage and in WAN optimization solutions. For data protection, deduplication is packaged as a feature of both software and hardware offerings. ESG's research found that adopters of deduplication selected a mix of hardware and software (see Figure 3); such fragmented deployment results in infrastructure management complexity, deduplication silos, and potentially higher costs.

Figure 3. Deduplication Implementation

**How is data deduplication technology implemented in your IT environment?
(Percent of respondents, N=140)**



Source: Enterprise Strategy Group, 2010.

In software, data can be deduplicated at the source production system via client agent technology. In this case, client software running on an application server identifies and transfers unique data to the backup media server and target storage device, providing greater network efficiency. Other software solutions deduplicate the backup stream at the backup server—removing any potential performance burden from production application servers. The deduplication domain is limited to data protected by the backup application; multiple backup applications in the same environment create deduplication silos.

Target-side deduplication typically leverages powerful purpose-built storage appliances to accommodate processing of the entire (non-deduplicated) backup load either pre- or post-ingestion. The interface for these systems varies (iSCSI or FC VTL, NAS, or [Symantec OpenStorage](#)). Often, the underlying architecture is not based on standard components. A typical deduplication target system is implemented and tightly integrated as part of the vendor's proprietary file system and software RAID stack—a more rigid implementation that makes it difficult to take advantage of new hardware platforms and capabilities, or deployment in different operating systems.

² Announced June, 2010.

HP's StoreOnce deduplication software has been designed as a modular component within the product architecture. It runs as an application/service in a standard Linux operating system, resides on a standard file system, and uses a standard RAID stack for data storage. This approach provides more architectural flexibility to extend the technology to different product deployments. It enables the software to be readily integrated with other components in HP's portfolio, including HP's scale-out file system and HP Data Protector backup software. It also means HP can leverage this technology as a component in a common storage software stack for HP Converged Infrastructure.

HP StoreOnce

HP StoreOnce deduplication software simplifies the deployment of deduplication technology across IT infrastructure. Not licensed as standalone software, it is a portable engine that can be easily embedded in multiple infrastructure components, eliminating the complexity seen in earlier-generation deduplication. HP StoreOnce uses patented innovation and features designed by HP Labs to maximize backup and recovery performance while minimizing management and hardware overhead.

HP StoreOnce deduplication software identifies replicate data inline (upon ingest) with its sparse index-based deduplication approach. This method has two phases:

1. HP StoreOnce algorithms sample large data sequences (approximately 10 MB) to identify the likelihood of duplicates and rapid routing delivers each sequence to the best node for deduplication.
2. StoreOnce uses a SHA-1 hash algorithm on approximately 4 KB variable-length blocks. By using a subset of key values stored in memory, StoreOnce determines a small number of sequences already stored on disk that are similar to any given input sequence. Then each input sequence is only deduplicated against those few sequences. This minimizes disk IO and uses less disk and little memory, creating more efficiency and enabling faster ingest and, importantly, restoration of data.

As previously mentioned, deduplication involves replacing duplicate data with pointers to existing (unique) data. If the unique data is scattered across a storage system (i.e., "fragmented"), then restoring it could take longer because reconstituting it would require many slow random seeks. StoreOnce avoids this situation by not replacing small amounts of duplicate data with pointers to faraway places with no other related data. This approach greatly improves restore speed with only a bit more extra data stored.

HP's approach has more far-reaching implications. The architecture and design of the deduplication software makes it portable, scalable, and able to deliver global deduplication (within and across independent multiple nodes with a single namespace). The implication is that HP StoreOnce deduplication can be deployed in a number of iterations; for example, as a virtual machine instance, integrated with HP Data Protector backup and recovery software, and with the HP X9000 scalable NAS storage. The architecture commonality also means these deployments can extend across the WAN and in ROBO environments without requiring data to be rehydrated and then deduplicated multiple times.

HP D2D Backup Systems

Initially, the 64-bit HP StoreOnce deduplication software is implemented in the HP D2D Backup System family of backup target storage solutions, is built on standard components, and is suitable for remote and branch offices (ROBO) up through mid-sized data centers with up to 36 TB of usable capacity:

- D2D2500 (1U): small environments and ROBOs with up to 3 TB of usable capacity.
- D2D4106 (2U): ROBOs and small to mid-sized data centers with up to 9 TB of usable capacity.
- D2D4112 (2U): ROBOs and small to mid-sized data centers with up to 18 TB of usable capacity.
- D2D4312 (4U): ROBOs and mid-sized data centers with up to 36 TB of usable capacity.³

³ With a 20:1 reduction ratio and retention period of more than 12 weeks, the logical capacity of the device is up to 720 TB.

HP D2D Backup Systems offer NAS, VTL, and Symantec OpenStorage (OST) interfaces to backup software. For NAS, one or more file shares can be created on each system which is used by the backup application as either CIFS or NFS targets for backup. The iSCSI or FC VTL interface emulates LTO tape autoloaders and libraries which the backup application sees as one or more target tape devices. For Symantec NetBackup and Backup Exec users, an OST interface can be used for the D2D appliance configured as a CIFS target. Instead of directing backups directly to CIFS shares on the D2D system, the media server is configured to use the OST plug-in (which resides on the NetBackup or Backup Exec media server). The interface enables backup catalog tracking of duplicated data between D2D devices.

For offsite disk copies, HP D2D device-to-device replication at the cartridge level (VTL) or share level (NAS) in 1:1 or many:1 (up to 50:1) configuration provides network-efficient transfer between devices. Bandwidth throttling restricts the amount of bandwidth being used during replication for even more network efficiency. Unique (duplicate or unmatched) data chunks are compressed before being stored to disk. HP D2D Replication Manager software is also available to make it easier to manage a large number of devices being replicated to a central site.

The D2D device enables longer retention on disk and device-to-device replication for DR copies, so it's less likely that data will need to be copied to physical tape. However, when periodic offload to physical tape for compliance or long-term retention is required, best practices suggest:

- **VTL mode:** Use the backup application to back up source data to physical tape as a separate activity or use the backup application to copy virtual tape cartridges (data is rehydrated) to physical tape cartridges in a physical device either on the SAN or attached directly to the media server.
- **NAS mode:** Use the backup application to back up source data to physical tape as a separate activity.

Deduplication Purchase Criteria

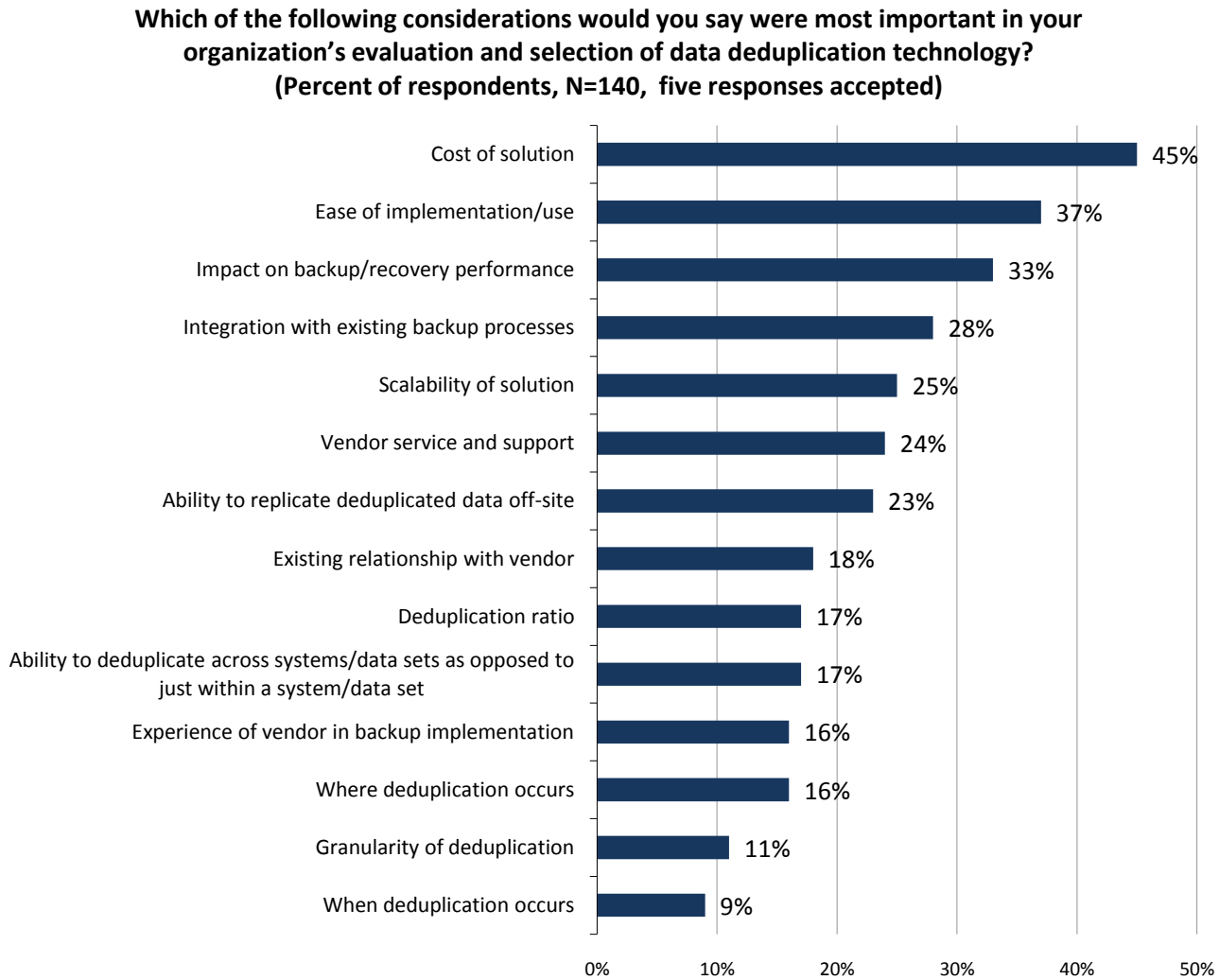
ESG research respondents ranked deduplication purchase criteria (see Figure 4). Examining HP StoreOnce deduplication software against these evaluation considerations shows several advantages of HP's approach.

Cost. Though deduplication can lower TCO in backup, cost is still the top concern when evaluating solutions. The weighting of cost should be balanced with how the solution mitigates risk and improves cycle time for operational staff managing backup and recovery operations. Deduplication delivers huge reductions in consumption, which would lower the impact and costs of data growth (hardware, bandwidth, and operational staff). HP claims price/performance leadership with its architecture; HP Converged Infrastructure offers volume economics and, when compared with leading deduplication solutions, HP D2D offers higher performance with less disk drives.

Ease of implementation and use. ESG research found that the ability to integrate with existing backup processes and overall ease of use are of greater importance to users than more specific technical considerations such as the deduplication ratio or the granularity of deduplication. HP StoreOnce deduplication is portable, modular, and will be packaged in hardware and software solutions. This more ubiquitous and standardized delivery of deduplication (multiple applications and data types, with easy to use management) provides flexibility and HP's application-agnostic design eliminates customization and complexity. In addition, the ability to package StoreOnce deduplication as a virtual appliance streamlines deployment and is more extensible to primary scale-out workloads.

Performance. Where and when deduplication happens could impact performance. It can occur at the client/source or the media server in software approaches and at the target device for hardware approaches. It can occur inline (before or during ingest) or post process (after all or a portion is written to disk). HP's improved inline performance comes from its efficient sparse index technology that matches data at high speeds with minimal memory requirements and from optimization that reduces the number of disk accesses required to process data.

Figure 4. Deduplication Purchase Criteria



Source: Enterprise Strategy Group, 2010.

Integration with existing backup solution. Hardware-based deduplication solutions are popular as they are easy to deploy and are often optimized for performance. Software approaches are popular because they are most-tightly integrated with backup processes (content-aware) and offer flexibility in disk storage selection. Solutions that strive to more tightly integrate hardware with software may deliver the best of both worlds.

Scalability. The ability to scale performance and capacity with ease and without disruption is key. HP StoreOnce technology is built on standard components, providing flexibility in product delivery and integration. It allows for the deployment of software as a virtual machine, integrated directly with backup software, or joined with other storage software assets. It has also been designed to support a distributed deployment in the future as part of a scalable multi-node storage system.

Ability to move copies off site. The most popular approaches are to replicate deduplicated data between local or remote devices over IP, and to create copies of “reinflated” data on physical tape media. While the recommendation is to make tapes under the knowledge and control of the backup software, HP offers both methods in addition to supporting Symantec OST for intelligent duplication. HP Data Protector will be aware of copies replicated by HP D2D devices.

Deduplication domain. With local deduplication, backup data is only deduplicated against other data processed within the same domain (which results in a silo storage approach to backup). Global deduplication allows backup data to be deduplicated against all other backup data (within and across systems receiving the data). One of the advantages of HP StoreOnce deduplication is that it is built upon an open architecture and its portability means that, in the future, HP will be able to marry it with the scale-out capability of HP's X9000 file systems to create scalable backup appliances for the data center. HP will be able to take advantage of the single namespace of the X9000 to deduplicate across physical systems.

HP Converged Infrastructure and HP StoreOnce

HP is talking to customers about its HP Converged Infrastructure, an overall architecture and delivery model that is akin to "liquid IT"—where dynamic delivery of resources and services is available. This concept takes advantage of IT resource pooling, seamless and dynamic provisioning, and ease of integration to provide on-demand delivery of IT services. The benefits are greater automation, dynamic resource utilization, planning ease, operational efficiency, and, ultimately, business agility. The delivery of the HP Converged Infrastructure includes design principles around virtual resource pooling and standard, modularized building blocks—concepts that align with what we have discussed in this paper relative to HP StoreOnce.

HP StoreOnce deduplication leverages HP Converged Infrastructure based on an extensive server, network, storage, and software technology portfolio. Instead of independent, siloed stacks for server, storage, and network resources, they're integrated. HP Converged Infrastructure allows IT organizations to focus more on delivering business value than on managing infrastructure operations.

The Bigger Truth

Data growth continues unabated. Early implementations of data deduplication have proven valuable in the battle against the effects of data growth. However, the haste of vendors to deliver deduplication solutions and of end-users to adopt them has led to inefficiency: scale-out limitations, reduction and re-inflation for data movement, and deduplication silos (especially between primary and secondary storage, hardware and software solutions). The fragmented nature of first generation deduplication has done little to eliminate sprawl—or the management overhead associated with it.

HP's vision is that data should be created, shared, and protected without necessitating alternating deduplication and "undeduplication" along the way. With these principles in mind, HP developed StoreOnce deduplication. StoreOnce deduplication technology's application-agnostic design eliminates customization and complexity. It is portable to backup workloads in virtual machines, simple appliances, data protection software, and scalable storage solutions and is extensible to primary scale-out workloads. Designed by HP Labs and built on HP Converged Infrastructure, StoreOnce offers high performance and efficiency and leverages volume economics, consistency across the data workflow, and management through a large single namespace.

While the issues resulting from data growth may force the implementation of deduplication solutions that have some shortcomings, IT organizations should be evaluating solutions with long-term impact for the next phase of growth in mind. HP StoreOnce deduplication software meets many of the key purchase criteria today and is expected to redefine considerations near-term. As organizations implement deduplication to address relentless data growth demands, HP's strategy and portfolio of deduplication solution should be afforded a closer look.



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

20 Asylum Street | Milford, MA 01757 | Tel:508.482.0188 Fax: 508.482.0128 | www.enterprisestrategygroup.com

4AA2-1223ENW