

HP StoreOnce: reinventing data deduplication

Reduce the impact of explosive data growth with HP StorageWorks D2D Backup Systems

Technical white paper

Table of contents

Executive summary.....	2
Introduction to data deduplication	2
What is data deduplication, and how does it work?	3
Customer benefits.....	4
Challenges with today's data deduplication	5
Introducing HP StoreOnce: A new generation of deduplication software	6
HP StorageWorks D2D Backup Systems, powered by HP StoreOnce	7
Benefits of HP D2D Backup Systems with HP StoreOnce	8
HP StoreOnce enabled replication	9
Why HP D2D with StoreOnce is different from other deduplication solutions	10
How does HP StoreOnce work? A deeper dive	11
Setting expectations on capacity and performance.....	12
HP StorageWorks VLS Backup Systems	12
HP StoreOnce: Tomorrow's view	13
For more information.....	14



Executive summary

Rather than spending IT dollars on infrastructure and overheads, today's businesses need their IT dollars to go toward delivering new applications to help them to be more competitive in the business they're in, help them enter new businesses, and to support business transformation. Businesses everywhere are counting on IT to deliver more value to their business. HP believes that IT has the resources to do it. The problem is, because of the sprawl of data and systems, those resources are tangled up in the overheads of legacy architectures and inflexible stacks of IT.

Data explosion is a primary contributor to IT sprawl, inefficiency, and waste. The data growth challenge is particularly acute when looking at how much time and money customers spend in managing data protection processes and the ever growing mountain of archival and offline data. The digital data universe grew to 800 billion gigabytes in 2009, an increase of 62 percent from the 2008 figures, and it doesn't stop there; the digital data universe is expected to grow by 44 times between 2010 and 2020.¹ It's no surprise then, that a 2010 Storage Priorities Survey² reported that Enterprise storage managers list two of their three top storage priorities to be related to data protection; namely data backup and disaster recovery. Data deduplication continues to be the likeliest new technology to be added to backup operations, with 61 percent of customers in the storage priorities survey either deploying or evaluating it, and that's on top of the 23 percent of current deduplication users.

Data deduplication has emerged as one of the fastest growing datacenter technologies in recent years. With the ability to decrease redundancy and so retain typically 20x more data on disk, it continues to attract tremendous interest in response to data growth. However, implementations of first generations of deduplication technology have become extremely complex, with numerous point solutions, rigid solution stacks, scalability limitations, and fragmented heterogeneous approaches. This complexity results in increased risk, extra cost, and management overhead.

This whitepaper introduces and describes StoreOnce software, next-generation deduplication technology from HP that allows for better management, higher performance, and more efficient data protection, while providing IT administrators with a cost-effective way to control unrelenting data growth.

Introduction to data deduplication

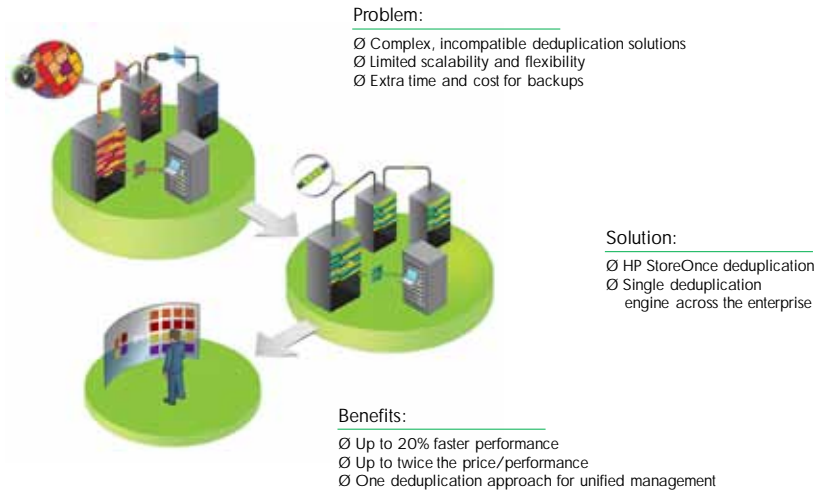
In recent years, disk-based backup appliances have become the backbone of a modern data protection strategy because they offer:

- Improved backup performance in a SAN environment with multi-streaming capabilities
- Faster single file restores than physical tape
- Seamless integration into an existing backup strategy, making them cost-effective and low risk
- The option to migrate data to physical tape for off-site disaster recovery or for long-term energy and cost-efficient archiving
- Optimized data storage through data deduplication technology, which enables more network efficient data replication

¹ IDC study "The Digital Universe Decade – Are You Ready?" May 2010.

² searchstorage.com March 2010.

Figure 1: HP is reinventing deduplication with HP StoreOnce

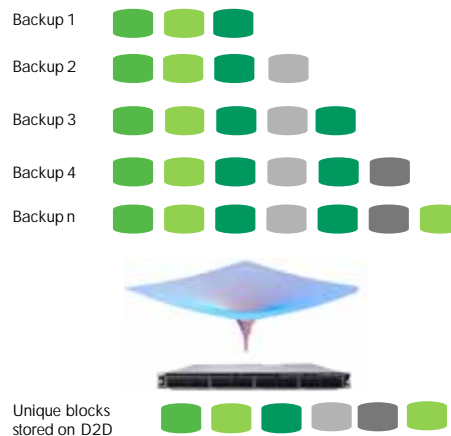


What is data deduplication, and how does it work?

Data deduplication is a method of reducing storage needs by reducing redundant data, so that over time only one unique instance of the data is actually retained on disk. Data deduplication works by examining the data stream as it arrives at the storage appliance, checking for blocks of data that are identical, and removing redundant copies. If duplicate data is found, a pointer is established to the original set of data as opposed to actually storing the duplicate blocks—removing or “de-duplicating” the redundant blocks from the volume (See Figure 2). However, indexing of all data is still retained so that it can be “rehydrated” should that data ever be required.

The key here is that the data deduplication is being done at the block level to remove far more redundant data than deduplication done at the file level (called single-instancing), where only duplicate files are removed.

Figure 2: Illustrating the deduplication process



As the backup process tends to generate a great deal of repetitive copies of data, and hence data deduplication is especially powerful when it is applied to backup data sets. The amount of redundancy depends on the type of data being backed up, the backup methodology and the length of time the data is retained.

Once the original file is stored, the technology removes duplicate data down to the block or byte level on all future changes to that file. If a change is made to the original file, then data deduplication saves only the block or blocks of data actually altered, (a block is usually quite small, less than 10 KB of data). For example, if the title of our 1 MB presentation is changed, data deduplication would save only the new title, usually in a 4 KB data block, with pointers back to the first iteration of the file. Thus, only 4 KB of new backup data is retained.

When used in conjunction with other methods of data reduction such as conventional data compression, data deduplication can cut data volume even further.

Customer benefits

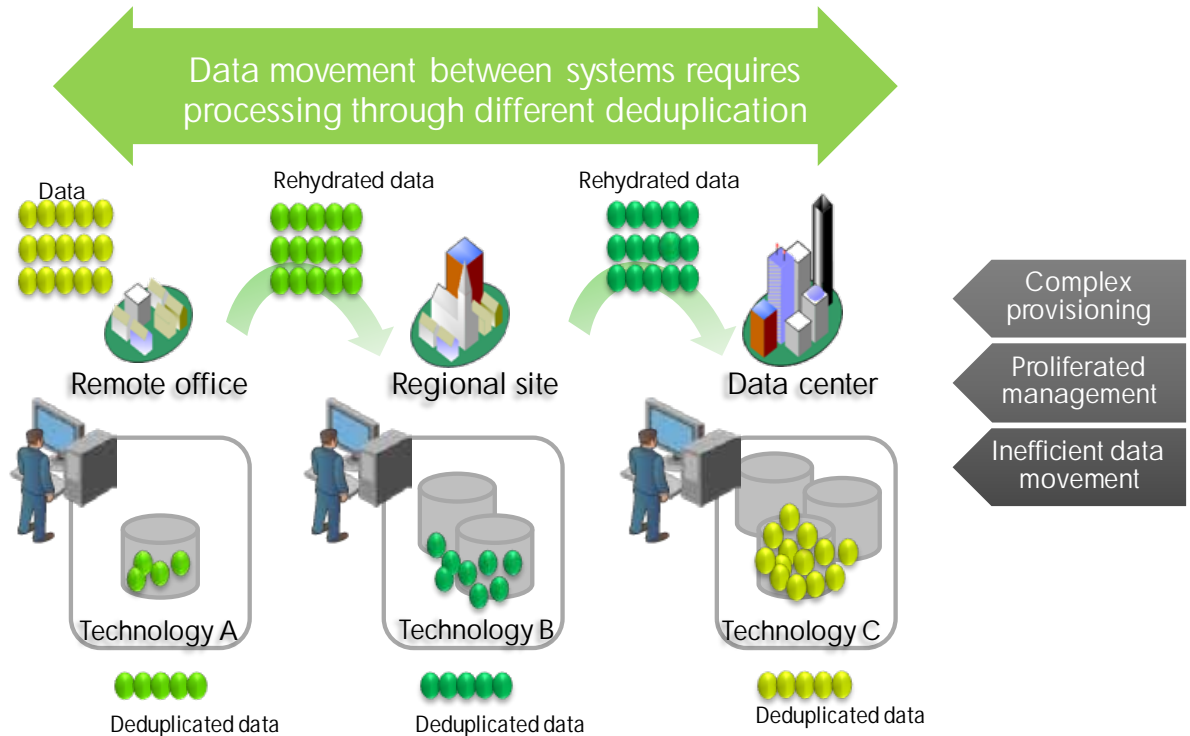
- **Ability to store dramatically more data online** (online means disk based): Retaining more backup data on disk for longer duration enables greater data accessibility for rapid restore of lost or corrupt files, and reduces impact on business productivity while providing savings in IT resource, physical space, and power requirements. Disk recovery of single files is faster than tape.
- **An increase in the range of Recovery Point Objectives (RPOs) available**: Data can be recovered from further back in time from earlier backup sets to better meet Service Level Agreements (SLAs).
- **A reduction of investment in physical tape**: This consequently reduces the overheads of tape management, by restricting the use of tape to more of deep archiving and disaster recovery usage model.
- **A network efficient way to replicate data offsite**: Deduplication can automate the disaster recovery process by providing the ability to perform site-to-site replication at a lower cost. Because deduplication knows what data has changed at the block or byte level, replication becomes more intelligent and it transfers only the changed data as opposed to the complete data set. This saves time and replication bandwidth, enabling better disaster tolerance without the need and operational costs associated with transporting data off-site on physical tape.



Challenges with today's data deduplication

First-generation deduplication has proven valuable in terms of data reduction; however, initial implementations have become extremely complex with point solutions, rigid solution stacks, scalability limitations, and fragmented heterogeneous approaches. This complexity results in increased risk, extra cost, and management overhead (as illustrated in Figure 3).

Figure 3: Today's data deduplication infrastructure—complex and fragmented



In this illustration (Figure 3), the data is deduplicated at the remote office, regional site, and data center. However, it must be rehydrated (also referred to as re-inflated or reduplicated) at each stage in order to be passed between incompatible deduplication technologies.

First-generation deduplication solutions were designed for either backup or primary storage, without really being optimized for both applications and suffered from the following issues:

Complex, incompatible backup solutions

- Fragmented, mixed-technology approaches to backup that have not been designed for a Converged Infrastructure that moves data and applications together
- Incompatible technologies in remote, regional, and central offices
- Multiple expansion and contraction cycles required when moving data from system to system, that is, data must be deduplicated and then rehydrated

Limited scalability and flexibility

- Lack of scalability across an enterprise, scaling is reliant on specific client-side backup software
- Tightly controlled, closed deduplication architectures

- Different optimizations for different application types, that is, solutions are not application agnostic and require a great deal of tuning for certain applications to function properly
- Rigid solution stacks that inhibit innovation and complicate new hardware deployments, for example, primary deduplication solutions are not effective across multiple isolated nodes

Extra time and cost for backups

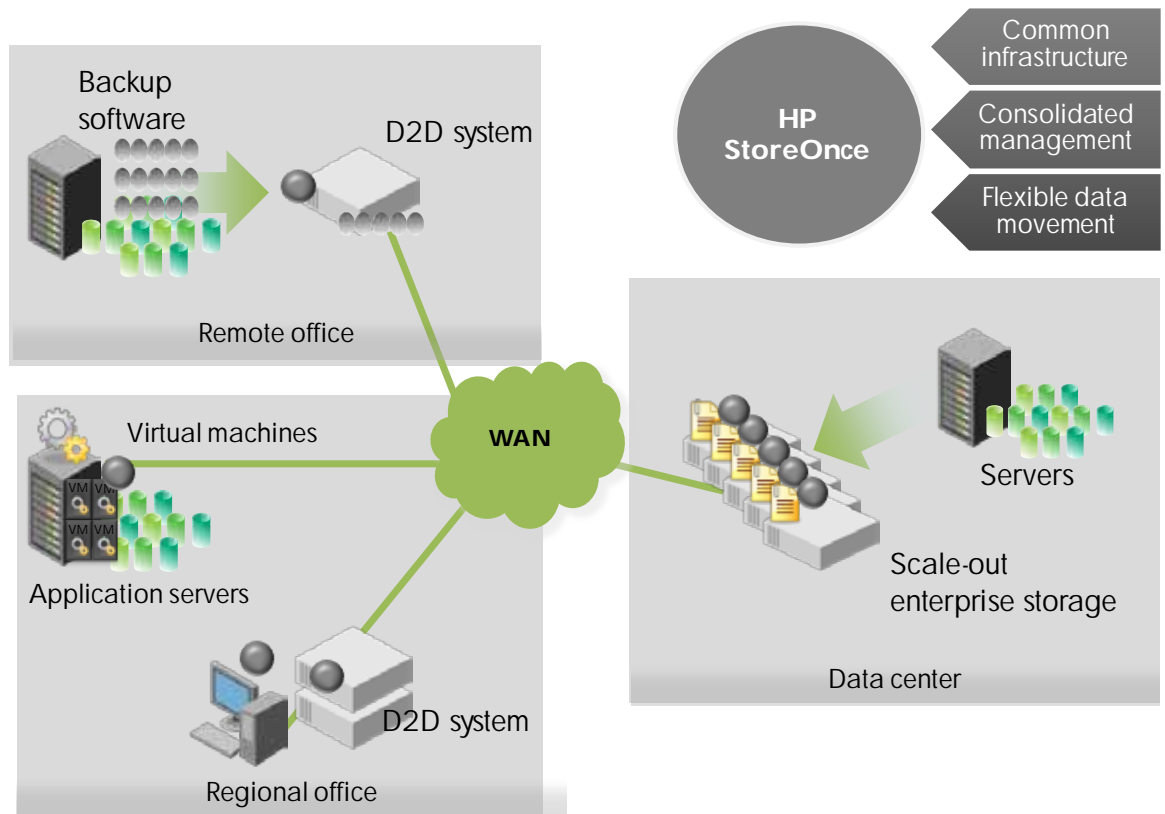
- Long backup and recovery times
- Multiple expansion/contraction cycles required when moving data from system to system
- More data sets to backup, more management burdens on IT administrators

It's our vision that data should be created, shared, and protected without being deduplicated and rehydrated (rededuplicated) multiple times along the way. Simply "StoreOnce" and that's it.

Introducing HP StoreOnce: A new generation of deduplication software

HP StoreOnce deduplication is a new class of software that leverages HP Labs technology innovations to deliver a more efficient way to accommodate data growth, without adding cost or reducing performance.

Figure 4: HP StoreOnce provides the same software solution from end-to-end.



HP StoreOnce deduplication can be used in multiple places in the data center. The advantage is that the same technology allows you to move/backup/access data without having to “rehydrate” the data before moving it to another point.

HP StoreOnce deduplication software delivers improved performance at half the price point of competitive solutions and enables clients to spend up to 95 percent less³ on storage capacity compared to traditional backup.⁴ HP StoreOnce software helps clients to enhance:

- **Performance**—Fast, efficient algorithms enable users to achieve 20 percent faster inline deduplication and twice the price/performance over competitive offerings through smart data and index layout capabilities that reduce disk utilization and increase input/output (I/O) efficiency.⁵
- **Scalability**—From small remote office to enterprise data center, the ability to support more storage for the same amount of CPU and memory. Simplify management and improve data protection by providing the ability to centrally run multiple backup sites and easily replicate data from remote offices to data centers.
- **Efficiency**—Smallest chunk size on the market means better, more resilient deduplication to data types and formats, as well as the ability to deduplicate data from adjacent use cases. This allows users to create, share, and protect their data without having to deduplicate multiple times.

HP StoreOnce software is available in all HP StorageWorks D2D Backup Systems from July 2010, providing a single technology that can be deployed at multiple points in a Converged Infrastructure. Previous generation D2D products soon may be able to upgrade to a 32 bit StoreOnce code stream allowing both generations to replicate between each other.

HP StorageWorks D2D Backup Systems, powered by HP StoreOnce

The HP StorageWorks D2D Backup Systems provide disk-based data protection for data centers and remote offices. Using HP D2D, IT or storage managers can automate and consolidate the backup of multiple servers onto a single, rack-mountable device while improving reliability by reducing errors caused by media handling. The D2D Backup Systems integrate seamlessly into existing IT environments and offer the flexibility of both NAS (CIFS/NFS) and Virtual Tape Library (VTL) targets⁶.

All HP D2D Backup Systems feature HP StoreOnce deduplication software for efficient, longer-term data retention on disk and enabling network-efficient replication for a cost-effective way of transmitting data offsite for disaster recovery purposes.

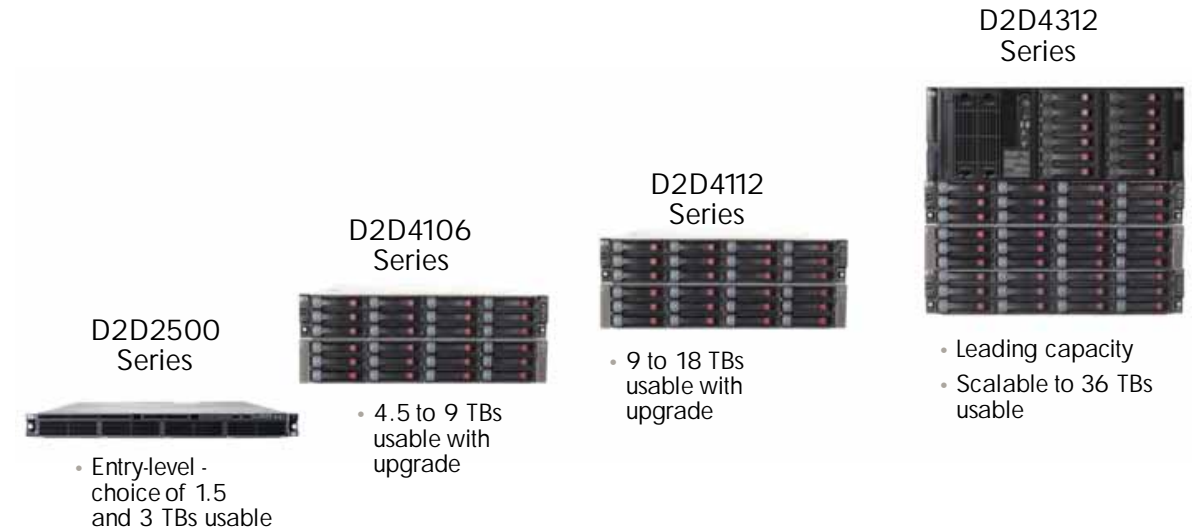
³ Based on multi-site deployment of D2D4312 ROI calculations as replacement technology for pure tape backup and offsite archival (HP D2D TCO Analysis from May 2010).

⁴ Based on average deduplication ratio of 20:1. Deduplication ratio could be as high as 50:1 under optimal circumstances.

⁵ Based on comparison to competitive systems using standard interfaces for all backup apps FC, CIFS, and NFS interfaces.

⁶ Earlier generations of D2D (before June 2010) only support NAS (CIFS) and virtual tape library (VTL) targets.

Figure 5: HP StorageWorks D2D Backup Systems



Benefits of HP D2D Backup Systems with HP StoreOnce

In addition to the benefits of disk-based backup, automation, and consolidation of multi-server backup to a single D2D appliance with HP StoreOnce deduplication, the HP StorageWorks D2D Backup Systems also provide the opportunity to:

Enhance business performance

- Get 20 percent faster in-line deduplication performance with innovations from HP Labs including smart data and index layout capabilities that reduce disk utilization and increase I/O efficiency.⁷
- Improve backup and recovery times with industry-leading in-line disk-based replication to improve business performance and reduce the impact of explosive data growth.
- Achieve throughput of up to 2.5 terabytes of data per hour and per node with HP D2D4312 Backup Systems.

Lower operational costs and maintain compliance

- Achieve up to twice the price/performance of competitive technologies—based on HP performance and pricing comparisons.
- Spend up to 95 percent less on new capacity⁸ and reinvest in business innovation.
- Store more backup data on disk for longer periods of time, and access data rapidly when needed for compliance purposes.
- Accelerate backup and recovery time while cutting your total business protection costs. Increase efficiency by allowing clients to create, share, and protect their data without having to “deduplicate” multiple times.
- Leverage a solution that outperforms competitive offerings in a real-world scenario.

⁷ Based on comparison to competitive systems using standard interfaces for all backup apps FC, CIFS, and NFS interfaces.

⁸ Based on multi-site deployment of D2D4312 ROI calculations as replacement technology for pure tape backup and offsite archival (HP D2D TCO Analysis from May 2010).

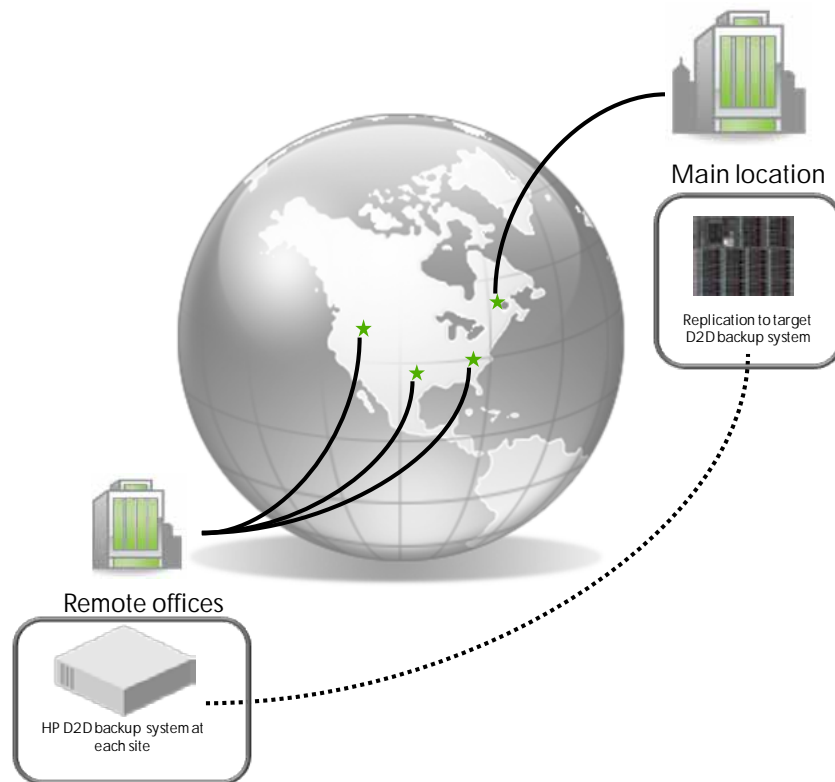
Improve business continuity and gain affordable data protection for ROBO sites

- Improve business continuity by providing cost-efficient replication of data across the storage network to an offsite location for disaster recovery purposes.
- Manage all local and remote backup with multi-site replication management.
- Simplify management overhead with a deduplication platform that's easy to deploy and use.
- Maintain remote systems cost-effectively with HP Integrated Lights-Out management (iLO) and integration with HP System Insight Manager.

HP StoreOnce enabled replication

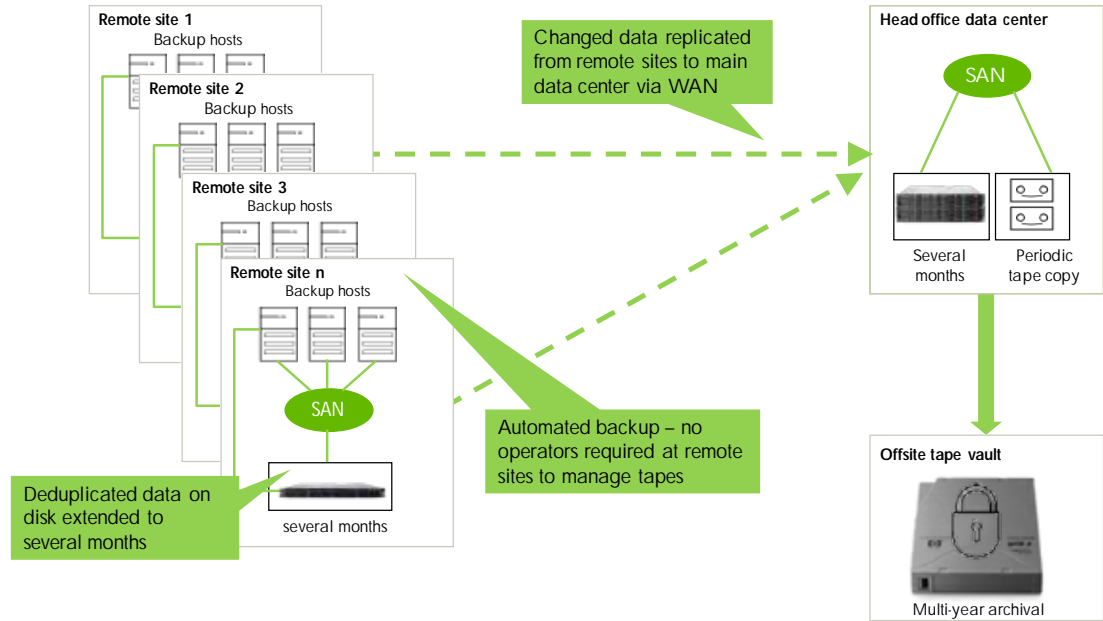
Data replication is the process of making a replica copy of a data set across a network to a “target site”. It is generally used to transmit backup data sets off-site to provide disaster recovery (DR) protection in the event of catastrophic data loss at the “source site.”

Figure 6: HP StoreOnce enabled data replication—affordable disaster recovery for ROBO sites



In the past, only large companies could afford to implement data replication as replicating large volumes of data backup over a typical WAN is expensive. However, because StoreOnce deduplication knows what data has changed at the block or byte level, replication becomes more intelligent and transfers only the changed data as opposed to the complete data set. This saves time and replication bandwidth making it possible to replicate data over lower bandwidth links for a more cost-effective, network-efficient replication solution. HP D2D Backup Systems with StoreOnce provide an automated and practical disaster recovery solution for a wide range of data centers, in addition to being an ideal solution for centralizing the backup of multiple remote offices.

Figure 7: Benefits of HP StoreOnce enabled replication



Data replication from HP features the ability to limit the replication bandwidth used for even more network-efficient replication. Without this ability, a replication job would use as much bandwidth as is available, potentially making other network activities unresponsive. Replication bandwidth limiting is customer-configurable at the appliance level via the graphical user interface and is set as a percentage of the available network bandwidth.

All HP D2D Backup Systems are available with an option to license data replication by target device. Note that HP D2D Replication Manager software is included with the license to provide an easier way to manage a large number of devices being replicated to a central site. HP OST plug-in installed on Media Servers and Symantec backup applications (i.e. Backup Exec or NetBackup) have the visibility to replicate copies of backups on remote D2D Backup Systems.

Why HP D2D with StoreOnce is different from other deduplication solutions

While competitors have unwittingly engineered complexity into their deduplication solution, HP has simplified the process and enabled 2x⁹ the price/performance of competing solutions. HP StoreOnce deduplication software includes technology innovations from HP Labs:

- HP StoreOnce software uses the smallest data block sizes in the industry to deliver application independence. This removes the need to maintain specific application optimizations and reduces application scaling issues that result from having multiple optimizations for different application types.

⁹ Evaluator Group independently tested and verified the HP 20% performance advantage and 2X price-performance claim. This was based on performance comparison to competitive systems using FC, VTL, and CIFS interfaces conducted in May 2010 and standard published pricing as of April 2010.

- Locality sampling and sparse indexing significantly lowers RAM requirements without sacrificing quality, it's like having a GPS for data deduplication.
- Intelligent data matching enables HP D2D to perform optimally for multiple data types and backup software deployments with no added complexity.
- Optimized data chunking results in minimal fragmentation compared to the competition, which reduces ongoing management overhead and speeds up restore time.

The result is a single unified architecture and a single deduplication software engine built on HP Converged Infrastructure. This extensible solution enables HP to deploy StoreOnce-based solutions from client to enterprise scale-out, enabling common management, and faster integration of new technologies.

How does HP StoreOnce work? A deeper dive

HP StoreOnce software works by using hash-based chunking techniques for data reduction. Hashing works by applying an algorithm to a specific chunk of data and yielding a unique fingerprint of that data. The backup stream is broken down into a series of chunks. For example, a 4K chunk in a data stream can be "hashed" so it is uniquely represented by a 20-byte hash code; a massive order of magnitude reduction.

The larger the chunks, the less chance there is of finding an identical chunk that generates the same hash code, and thus the deduplication ratio may not be as high. The smaller the chunk size, the more efficient the data deduplication processes but greater indexing overhead.

HP StoreOnce follows a straightforward process:

1. As the backup data enters the target device (in this case the HP D2D2500, D2D41XX, or D2D4300 Series Backup Systems); it is chunked into a series of average 4K size chunks against which the SHA-1 hashing algorithm is run. Batches of thousands of these chunks are deduplicated at a time.
2. A small number of chunks are sampled from each batch and then looked up (via their hash) in a small in-RAM sparse index. The index maps each sampled chunk to previously backed up sections of data containing that chunk. Because the index needs to map only the sampled chunks, not all chunks, it can be two orders of magnitude smaller than some alternative technologies.
3. Using the results from the index, HP StoreOnce chooses a small number of previously backed up sections that are highly similar to the new batch of data and deduplicates the new batch against only those sections. It does this by loading hash lists of the chunks in each section and comparing to see if any of them match the hashes of the chunks in the new batch.
4. Chunks without matching hashes are written to the deduplication store and their hashes are also stored as entries in a recipe file. The recipe represents the backup stream and it points to the locations in the deduplication store where the original chunks are stored. This happens in real time as the backup is taking place. If the data is unique (that is, a first time store), the HP D2D uses the Lempel-Ziv (LZ) data compression algorithm to compress the data after the deduplication process, and before storing it to disk. Typical compression ratios are between 1.5:1 and 2:1, but may be higher depending on the data type.
5. Chunks whose hash matches duplicate are not written to the deduplication store. An entry with the chunk's hash value is simply added to the "recipe file" for that backup stream pointing to the previously stored data, so space is saved. As you scale this up over many backups, there are many instances of the same hash value being generated. However, the actual data is only stored once, so the space savings increase.

6. To restore data from the backup system, the D2D device selects the correct recipe file and starts sequentially re-assembling the file to restore. It reads the recipe file, getting original chunks from disk using the information in the recipe, and returns the data to the restore stream.
7. Chunks are kept until no recipe refers to them (for example, the backups containing them have been deleted), then they are automatically removed in a housekeeping operation.

Most deduplication technologies in existence today use hash-based chunking. However, they often face issues with the growth of indexes and the amount of RAM storage required to store them. For example, a 1 TB backup data stream using 4K chunks results in 250 million 20-byte hash values requiring 5 GB of storage. If the index management is not efficient, this can significantly slow the backup down to unacceptable levels or require a great deal of expensive RAM.

By comparison, HP StoreOnce with HP Labs innovation dramatically reduces the amount of memory required for managing the index without noticeably sacrificing performance or deduplication efficiency. Not only does this technology enable low-cost high performance disk backup systems, but it also allows the use of smaller chunk sizes. This produces more effective data deduplication that is more robust to variations in backup stream formats or data types.

Setting expectations on capacity and performance

In terms of how much more data can be stored with HP StoreOnce and D2D Backup Systems, the simple answer is that you can expect to see a reduction in the amount of data typically by 20x. However, this is dependent on the nature of the data being backed up, the backup methodology, and the length of time that the data is retained. With the backup of a typical business data set over a period of more than 3 months, users might expect to see a deduplication ratio of 20:1 for daily incremental and weekly full backups, or more for daily full backups.

Performance is highly dependent upon specific environments, configuration, type of data, and the HP D2D Backup System model that is chosen.

For more detailed and specific advice, we recommend that you use the HP sizer tool at: www.hp.com/go/storageworks/sizer

A best practices whitepaper can also help with performance tuning for the HP D2D Backup System—refer to: <http://h20195.www2.hp.com/v2/getdocument.aspx?docname=4AA2-7710ENW.pdf>

HP StorageWorks VLS Backup Systems

The VLS product family with its accelerated deduplication capabilities continues to be an important part of the portfolio for large enterprise customers backing up to disk over a Fibre Channel SAN. HP is continuing to invest in this solution with improvements in performance and capacity. That said, the D2D product family with its improvements in performance and scalability aims to be the primary solution for customers backing up over Ethernet via VTL, NAS, or OST with a common architecture from client to back-end and on both physical and virtual devices.

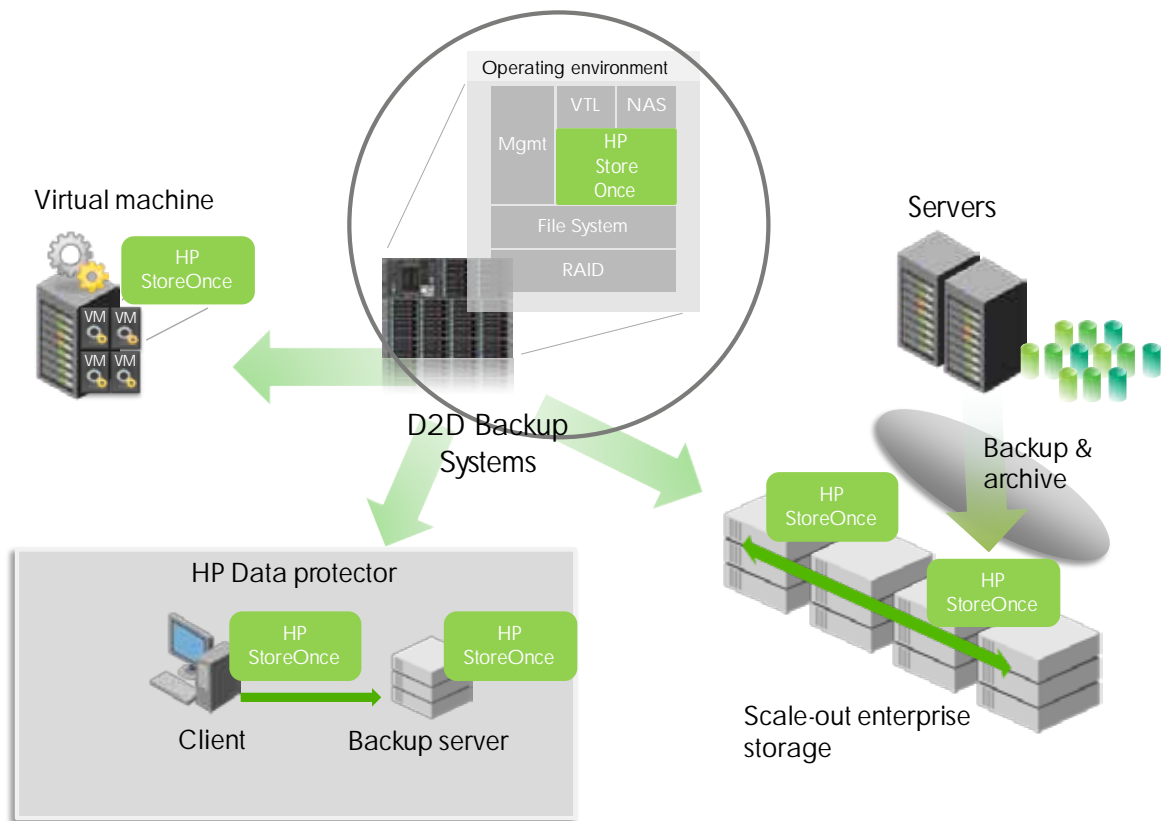
HP StoreOnce: Tomorrow's view

HP StorageWorks continue to help clients build a Converged Infrastructure with virtualized storage solutions that reduce IT sprawl and provide the operational flexibility to shift resources easily for improved business productivity. HP StorageWorks is transforming data protection with HP StoreOnce, the next generation in deduplication software designed to provide unprecedented performance, simplicity, and efficiency while maintaining business continuity.

Due to its modular design, future HP StoreOnce software can be extended across a number of solutions addressing different environments and deployments. One of these is as a virtual machine appliance providing flexibility to address environments too small for a dedicated appliance. Another is integrated with HP Data Protector Software, both as a software deduplication target store, as well as an option to employ deduplication on the client, reducing data sent over the network to the source. Because HP StoreOnce has been designed to support distribution of deduplication across a multi-node system, it can be integrated with the HP file system technology in the X9000 to provide a scale-out storage system for the enterprise for backup and archive data.

Through all this, the HP StoreOnce portfolio maintains compatibility between the various implementations. This allows customers to mix and match StoreOnce solutions, whether hardware versus software, target versus client-side, or remote office versus enterprise implementations. The end result enables a major step towards simplified management and flexible data movement across the enterprise.

Figure 8: Extending the technology—modular architecture enables flexible deployment and integration



For more information

HP D2D central:

www.hp.com/go/d2d

www.hp.com/go/storeonce

Sizing and configuration advice:

www.hp.com/go/storageworks/sizer

<http://h20195.www2.hp.com/v2/getdocument.aspx?docname=4AA2-7710ENW.pdf>

To know more about how deduplication with StoreOnce can help you store data effectively with higher data protection in a cost-effective way, visit www.hp.com/go/d2d

Share with colleagues



© Copyright 2008-2010 Hewlett-Packard Development Company, L.P. The information contained herein is subject to change without notice. The only warranties for HP products and services are set forth in the express warranty statements accompanying such products and services. Nothing herein should be construed as constituting an additional warranty. HP shall not be liable for technical or editorial errors or omissions contained herein.

4AA1-9796ENW, Created May 2008; Updated July 2010, Rev. 2

