Integrating HP Data Protector software with HP Data Deduplication Solutions

An analysis of how to implement data deduplication technologies utilizing HP Data Protector software

Executive summary

This white paper provides complementary information on data deduplication technologies supported by the latest Storage Solutions from HP. Data deduplication is a hot topic in data protection and therefore, also a relevant topic for HP Data Protector software.

Solution description

HP Data Protector software is a backup and disaster recovery product that provides reliable data protection and high availability for your expanding mission critical data. HP Data Protector network component concept provides for tailor-made backup and recovery solutions ranging from a single system to thousands of systems across multiple sites. HP Data Protector software fully supports HP data deduplication technologies allowing you to recover files more quickly while reducing your data management and storage costs. Data deduplication can increase your storage efficiency by a ratio of 50:1—that's up to 5000%! The extra capacity allows you to keep more backup data online and ready to restore at a moment's notice. Overall, the increase in storage efficiency brought about by deduplication lets you do more for less.

What is data deduplication?

Data deduplication is the ability of an appliance or software to compare blocks of data being written to the backup device with data blocks previously stored on the device. If duplicate data is found, a pointer is established to the original data, rather than storing the duplicate data sets. This removes, or "deduplicates," the redundant data blocks. Data deduplication is done at the block or chunk level, not at the file level.

This greatly reduces the volume of data stored.

Data deduplication is often used in conjunction with other forms of data reduction, such as conventional data compression, to further reduce the data volume stored.

The best approach to data deduplication depends on your size and backup needs.

- **Deduplication for enterprises:** Object-level differencing, or *accelerated deduplication*, is a good choice for enterprise customers because it focuses on performance and scalability. It delivers the fastest restores, as well as the fastest possible backup by deduplicating data after it has been written to disk. You can scale up to increase performance simply by adding extra nodes.

- **Deduplication for midsize businesses and remote enterprise sites:** Hash-based chunking, or *dynamic deduplication*, is a good choice for small and midsize businesses or large enterprises with remote sites because it focuses on compatibility and cost. It delivers a low-cost, small footprint in a format-independent solution.

A detailed description about deduplication techniques can be found in the "Understanding the HP Data Deduplication Strategy" HP white paper at : http://h71028.www7.hp.com/ERC/downloads/4AA1-9796ENW.pdf

Figure 1 shows the principal deduplication concept.

Figure 1: Deduplication Concept
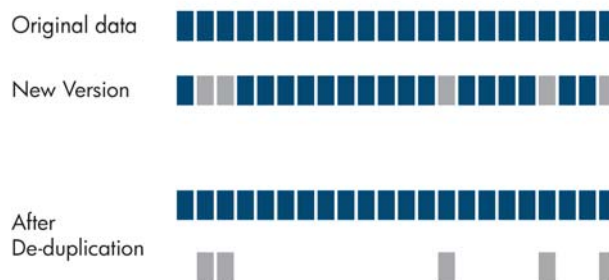
The benefits of data deduplication

There are a number of benefits realized by using data deduplication technology. The most compelling one is the increase of effective capacity for storing backup data. This allows for longer retention periods for backup data on disk, resulting in faster data recovery and higher service level agreements. HP disk-based backup systems with built-in data deduplication also reduce space and power requirements for the increased volume of protected data.

Read more about the deduplication benefits in the "HP Dynamic Deduplication—achieving a 50:1 ratio" HP white paper at: http://h71028.www7.hp.com/ERC/downloads/4AA2-0212ENW.pdf

Details to know about object-level differencing

Object-level differencing (accelerated deduplication) provides the best performance as the deduplication of backup data is a post-backup process. That is the reason why the backup device (Virtual Tape Library) has to be knowledgeable in terms of backup formats and data types to understand the meta data. HP Accelerated deduplication will support HP Data Protector 6.0 software and specific data types at launch.

Initially by HP Accelerated deduplication supported data types are:

- File system backups
- RAW Disk
- Microsoft Exchange

More data types and future Data Protector software versions will be added to the support matrix over time. The "HP StorageWorks Enterprise Backup Solution (EBS) Hardware/Software Compatibility Matrix" can be found at: http://www.hp.com/go/ebs.

How much space does data deduplication really save?

Two significant factors affecting the deduplication ratio for backups are:

- How long do you retain the data
- How much does data change between backups

Example: 500 GB file server backup

Retention Policy

- 1 week, daily incrementals (5)
- 6 months, weekly fulls (25)

Data parameters

- Daily change rate = 1% (10% of data in 10% of files)
- No compression

	Data stored normally	Data stored with deduplication
1st daily full backup	500 GB	500 GB
1st daily incremental backup	50 GB	5 GB
2nd daily incremental backup	50 GB	5 GB
3rd daily incremental backup	50 GB	5 GB
4th daily incremental backup	50 GB	5 GB
5th daily incremental backup	50 GB	5 GB
2nd weekly full backup	500 GB	25 GB
3rd weekly full backup	500 GB	25 GB
25th weekly full backup	500 GB	25 GB
Total	**12,750 GB**	**1,125 GB**

This example uses a system containing 500 GB of backup data that equates to 500 GB of storage for the first traditional full backup. If 10% of the files change between backups, then a traditional incremental backup would send about 10% of the size of the full backup or about 50 GB to the backup device. However, because data deduplication operates at the block level, instead of the file level, in actuality only a 1% change in the data has occurred. This means only 5 GB of block level changes or 5 GB of data stored with deduplication. Over time, the savings multiply. When the next full backup is stored, it will not be 500 GB. With deduplication the equivalent full backup is only 25 GB. A backup system with data deduplication enabled would use the same amount of storage in six months that would typically be required to store only one week of traditional backup data. Over a 6 month period data deduplication would provide an 11:1 effective savings in storage capacity. It also provides the ability to restore from further back in time without having to go to physical tape for the data. The key thing to remember here is that the deduplication ratio depends primarily on two things:

- What percentage of the data is changing between backups (percentage of data in percentage of files)
- How long is the retention period of the backups stored on disk

For example, a 0.5% daily change in the data in 10% of the files would yield a 50:1 deduplication ratio over one year of daily full backups. Obviously, the percentage daily change rate is quite difficult to predict for complex systems, especially for applications like Exchange, SQL, and Oracle so benchmarking is strongly advised.

As already indicated, backup data retention period and backup data change rate matters to find out what the approximate deduplication ratio will be. Figure 3 shows the approximate space saving based on the given daily change rate and backup policy.

Daily change rate	Backup policy					
	Daily full and weekly full			Daily incremental and weekly full		
	4 months*	6 months	1 year	4 months*	6 months	1 year
0.50%	31:1	37:1	50:1	25:1	31:1	46:1
1.00%	24:1	27:1	32:1	19:1	23:1	29:1
2.00%	16:1	17:1	18:1	13:1	15:1	17:1
* 4 months = 5 daily + 17 weekly backups Ratio = data sent vs. data stored						

Deduplication portfolio strategy from HP

HP has selected two deduplication technologies—one for enterprises and one for the SME and remote offices.
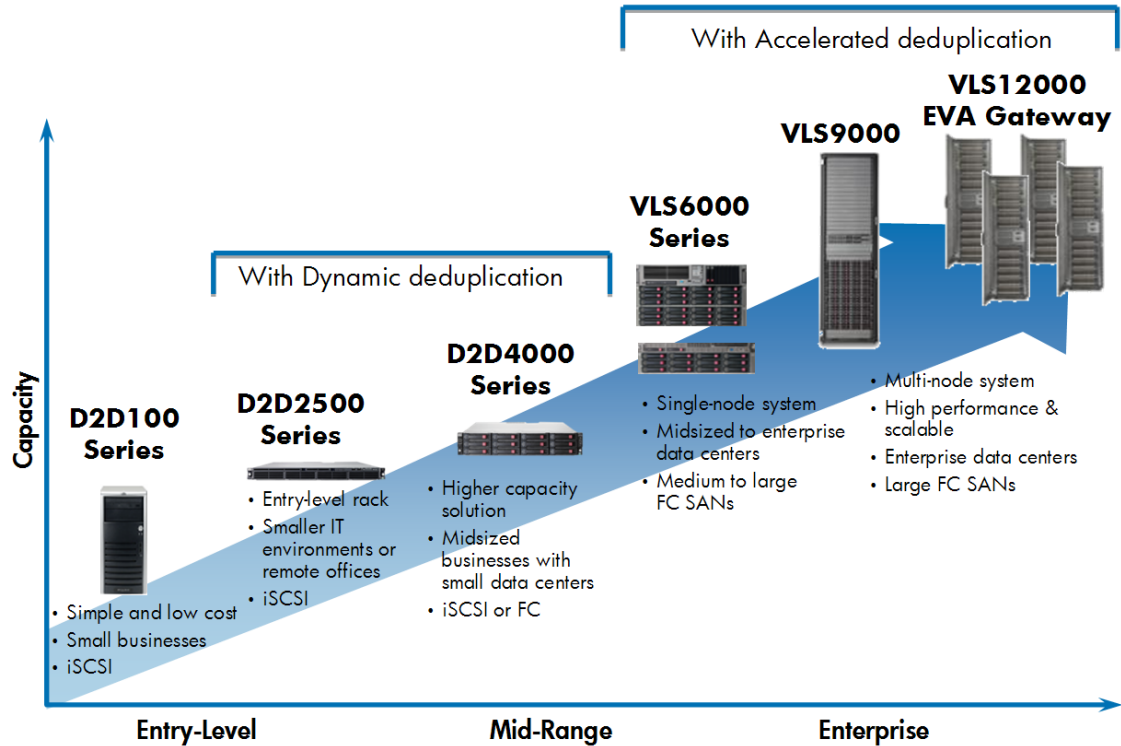
- Accelerated deduplication, available for HP StorageWorks Virtual Library Systems.
  Used on **HP VLS6000/9000/12000**
- Dynamic deduplication, built into HP StorageWorks D2D Backup Systems.
  Used on **HP D2D2500** and **D2D4000**

Figure 4 depicts the latest announced Entry-to-Enterprise level storage devices and their deduplication capabilities.

The HP StorageWorks D2D2500 and D2D4000 Backup Systems implement HP Dynamic deduplication technology. These range in size from 2.25 TB to 7.5 TB and are aimed at remote offices or small enterprise customers. The D2D2500 has an iSCSI interface to reduce the cost of implementation at remote offices, while the D2D4000 offers a choice of iSCSI or 4 Gb FC.

The HP StorageWorks Virtual Library Systems implement HP Accelerated deduplication technology and are all 4 Gb SAN-attach devices that range in native user capacity from 4.4 TB to over a petabyte with the VLS9000 and VLS12000 EVA gateway. Hardware compression is available on the VLS6000, 9000, and 12000 models, achieving even higher capacities. The VLS9000 and VLS12000 use a multi-node architecture that allows the performance to scale in a linear fashion. With eight nodes, these devices can sustain a throughput of up to 4800 MB/sec at 2:1 data compression, providing the SAN hosts can supply data at this rate. HP Virtual Library Systems will deploy the HP Accelerated deduplication technology.

Figure 4: HP StorageWorks Disk-based Backup

Where does HP Data Protector software fit into the picture?

Deduplication, today, is a feature provided by HP StorageWorks Virtual Library Systems and HP StorageWorks D2D Backup Systems and other vendors' hardware.

Deduplication on the in the previously mentioned paragraph storage devices happens either as "Inline deduplication" (Dynamic deduplication) or as "Post-Process" (Accelerated deduplication) technique.

Both ways are completely transparent to HP Data Protector software.

HP Data Protector Advanced Backup to Disk Licensing

As of 1st July, 2008, the HP Data Protector Advanced Backup to Disk license will cover the planned/consumed capacity on HP Data Protector file libraries and virtual tape libraries (VTL).

In cases where HP Data Protector software is using the VTL exclusively, it is recommended to license a quantity of Advanced Backup to Disk licenses matching the physical capacity of the VTL. HP calls the physical VTL capacity "usable native capacity." Other vendors call it "raw capacity." The new physical size/consumption licensing model does not require compression rates and deduplication ratios to be considered. RAID overhead does not to be considered as well.

The relevant HP Data Protector Advanced Backup to Disk licenses are:

- B7038AA capacity license for 1 TB of backup disk storage
- B7038BA capacity license for 10 TB of backup disk storage
- B7038CA capacity license for 100 TB of backup disk storage

Verify capacity for VTL

The recommended tool to verify consumed or allocated disk space on virtual tape libraries is the web-based
Command View VLS management interface.

Figure 5 depicts the relevant view within the Command View VLS management interface to evaluate the
amount of allocated and consumed disk space on the VTL.
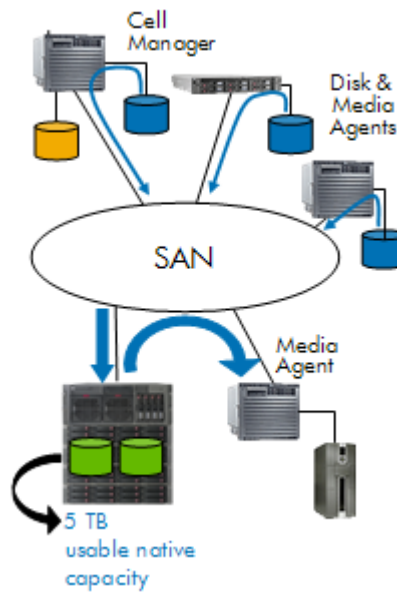
Figure 5: VLS Management Interface
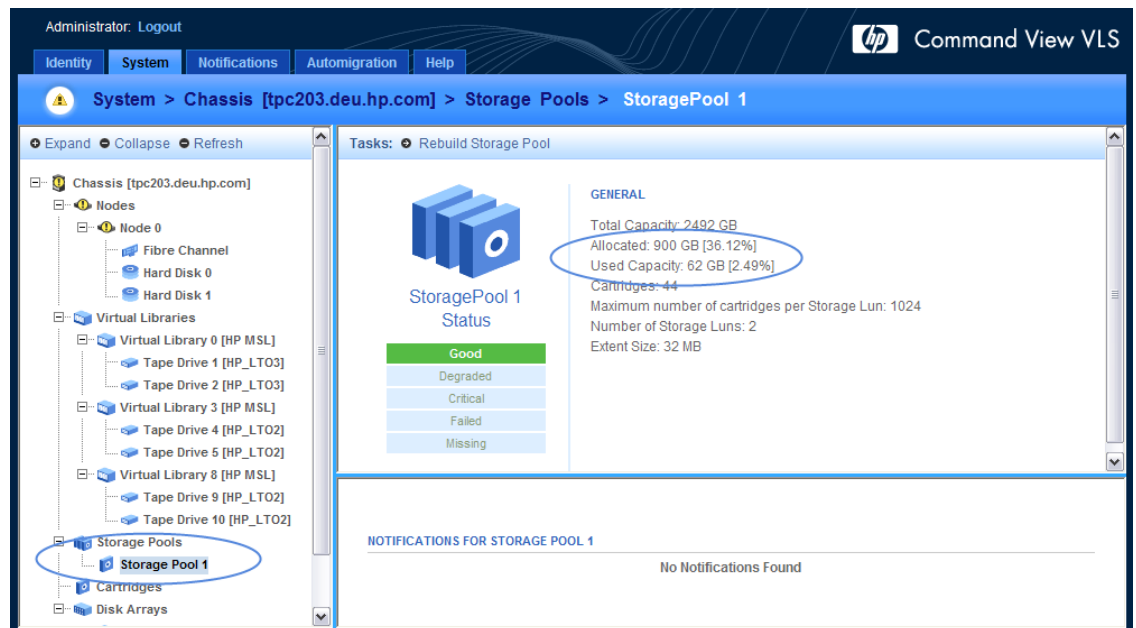
Licensing example

VTL configuration

In the example depicted in Figure 6, the virtual tape library (VTL) holds 5 TB of protected backup data
managed by HP Data protector software.

Figure 6: VLS Licensing Example

Usable native capacity of a virtual tape library (VTL) is the size on disk of the virtual tape library consumed by all protected HP Data Protector backups as reported by the VTL.

A total of five B7038AA Advanced Backup to Disk licenses are required to allow HP Data Protector software to utilize all 5TB in the example depicted in Figure 6.

Note

Virtual Tape Libraries can be upgraded using capacity-kits to extend the VTL's usable native capacity. Further backups to the virtual tape library (VTL) in the example in Figure 6 will exceed the total licensed Advanced Backup to Disk capacity therefore additional licenses will be required. Proper capacity planning is mandatory to not exceed the total licensed capacity.

For more information

Demystifying data deduplication:
http://h71028.www7.hp.com/ERC/cache/49205-0-0-0-121.aspx?bodycontentparams=583627-0-0-0-121&ERL=true

Assessing Data Deduplication Efficiency:
http://h71028.www7.hp.com/ERC/cache/49205-0-0-0-121.aspx?bodycontentparams=587598-0-0-0-121&ERL=true

HP StorageWorks D2D Backup Systems (D2D): (http://www.hp.com/go/d2d).

HP StorageWorks Virtual Library Systems (VLS): (http://www.hp.com/go/vls).

Data Protection Solutions with Deduplication: (http://www.hp.com/go/deduplication)

Technology for better business outcomes