Deduplication

PowerVault DL Backup to Disk Appliance

symantec™

Executive Summary

Introduction

Customers of all sizes and needs are seeking new ways to tackle their data protection challenges. While the challenges of data growth are not new, the pace of growth has become more rapid, the location of data more dispersed, and linkages between data sets more complex. Data deduplication offers companies the opportunity to dramatically reduce the amount of storage required for backups and to more efficiently centralize backup data from multiple sites for assured disaster recovery. The PowerVault DL Backup to Disk Appliance powered by Symantec Backup Exec 2010 now includes integrated deduplication capabilities that can help with the task of managing these data protection challenges. Deduplication capabilities are available when customers purchase the Deduplication Option.

What is Deduplication?

What is deduplication? At the core, deduplication is a process that breaks down files and data into “segments” and uses a tracking database to ensure the Media Server only stores a single copy of that segment across all client backup data stored to that media server. For subsequent backups of any client, the tracking database knows what segments have been protected and only transfers and stores the segments that are new or unique – file segments that are not currently stored by that Media Server. For example, if five different client systems are backing up data to a PowerVault DL Backup to Disk Appliance and a file segment is found that exists on all five of those client systems, only a single copy of the segment will actually be stored by the PowerVault DL Backup to Disk Appliance. This tracking database ensures that these segments are kept until any existing disk-based backup no longer references them. Because only a fraction of the original data is eligible to be stored by the PowerVault DL Backup to Disk Appliance, this leads to significant reduction in disk space needed for backups.

Backup Exec’s deduplication technology will deduplicate data across all servers that are being protected by the PowerVault DL Backup to Disk Appliance. The benefit of this methodology is that all of the deduplication segment information mentioned above is shared with all other backups configured to use deduplication for a specific PowerVault DL Backup to Disk Appliance. For example, if two Windows 2008 R2 servers are protected using either Client or Media Server deduplication, only deduplication segments that are unique to either of those servers will be stored. This helps significantly reduce backup disk utilization across all local and remote servers protected with deduplication.

Regardless of the methodology used for deduplication – Client or Media Server – the end result is the same: storage is optimized by only storing unique parts of a particular file or data stream, and using some form of a database to associate segments to each other and to the machines where they were backed up from.

With the Backup Exec 2010 Deduplication Option, Administrators have the ability to choose when and where deduplication takes place. Administrators can mix and match deduplication types to fit their unique needs; for example, a single media server licensed with the Deduplication Option can simultaneously use Client Deduplication for some jobs and Media Server deduplication for others.

- Client Deduplication is a software-driven process, where deduplication takes place at the **SOURCE** of data and is sent over the network in deduplicated form
- Media Server Deduplication is a software-driven process, where deduplication takes place **JUST BEFORE THE DATA IS STORED TO DISK** (also known as Inline Deduplication)

Each approach has its benefits and will be detailed in turn in the following sections.

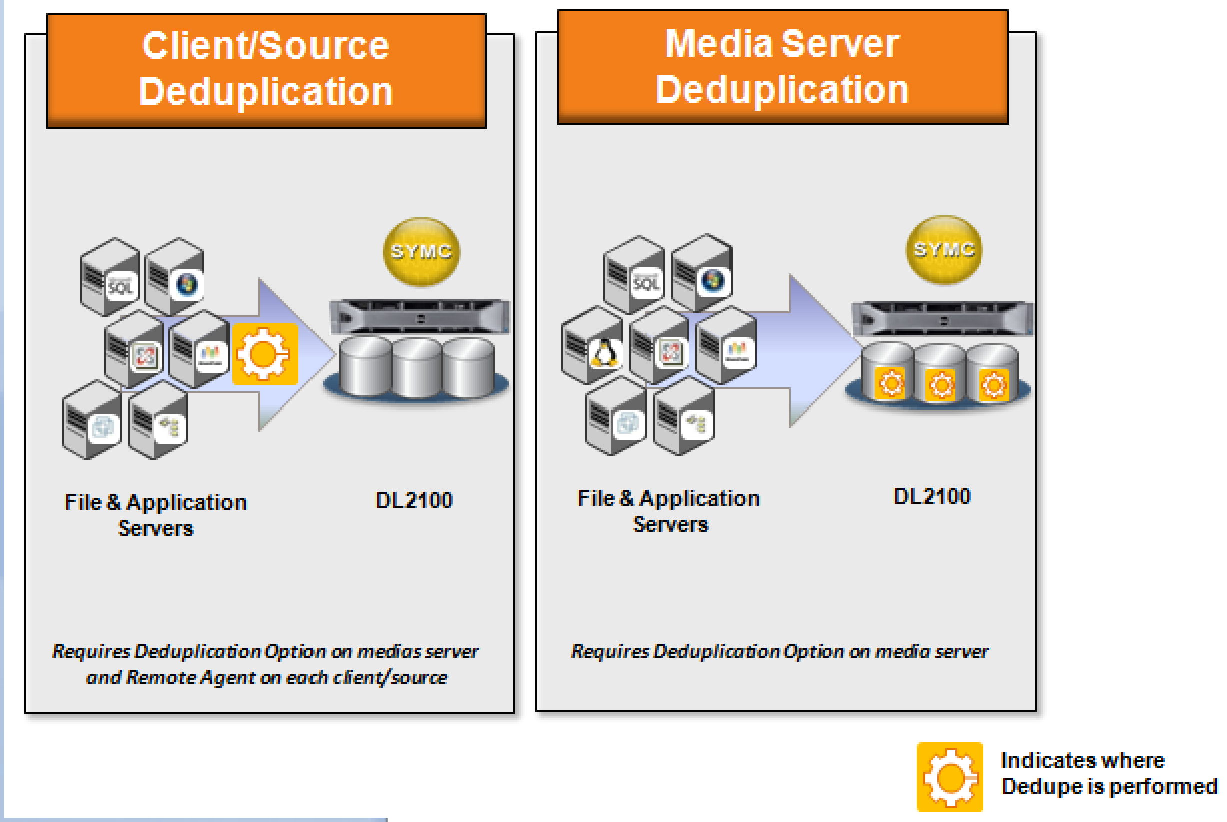

Client/Source Deduplication
SYMC
File & Application Servers
DL2100
Requires Deduplication Option on medias server and Remote Agent on each client/source
Media Server Deduplication
SYMC
File & Application Servers
DL2100
Requires Deduplication Option on media server
Indicates where Dedupe is performed

Figure 1: Deduplication Types using the PowerVault DL Backup to Disk Appliance.

Client Deduplication

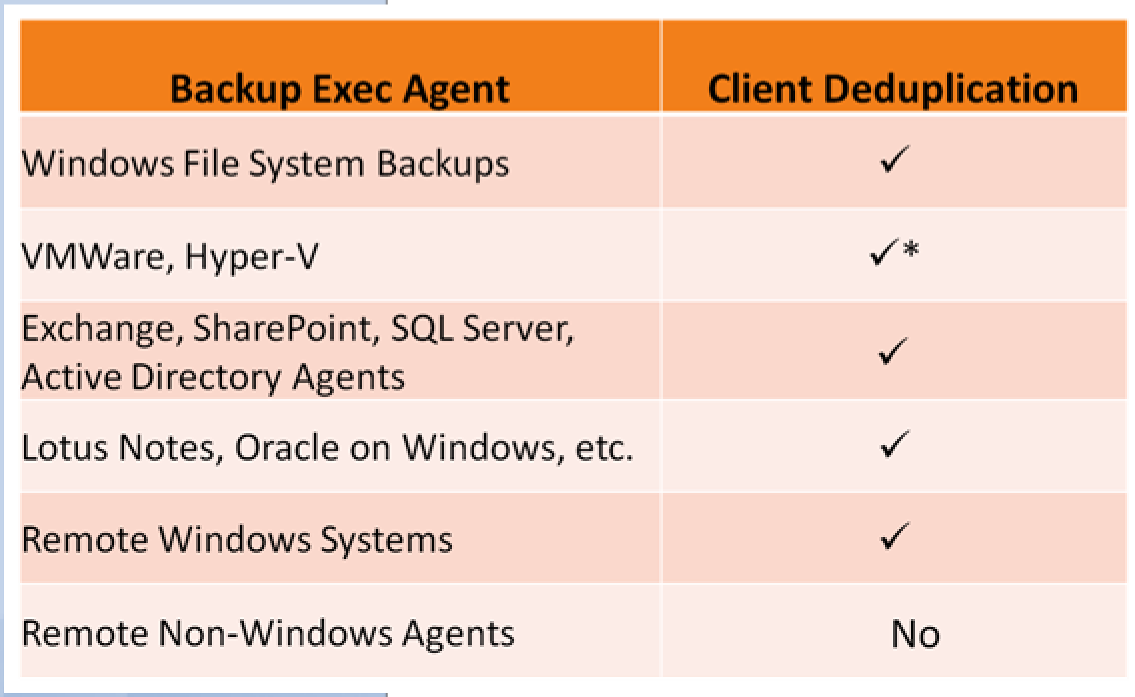
With Backup Exec 2010, exciting new possibilities for remote office protection were introduced. The concept of client deduplication – where the remote system is responsible for deduplication calculations and backup data is sent over the network in its deduplicated form – can make the process of protecting remote offices a much more streamlined experience. Remote offices can be challenging to protect effectively; WAN environments can be unreliable and can only utilize a fraction of the bandwidth available to LAN backups. Another common scenario exists where the customer’s VMware Server or Exchange Servers are the most powerful machines available in terms of processor speed or disk throughout.. Where appropriate, why not leverage some of this remote computing power to achieve faster backups? Both of these situations are examples where client deduplication can offer a comprehensive solution to the data protection challenges brought on by the environment.

Generally, remote office backup strategies have two basic architectures. First, there are remote offices which do not have local storage, and backup data is sent directly over the LAN or WAN to the data center for protection. Second, there are remote offices that employ local storage and then “forward” that locally stored backup data to the data center for protection. Both of these configurations can use Backup Exec 2010’s Deduplication Option to streamline and improve backup and recovery for remote offices.

Client deduplication is the default deduplication method Symantec recommends for several reasons:

- **Client deduplication enables greater scalability:** Client deduplication spreads processor usage out across all clients running backups, enabling the Media Server to process more concurrent backups.
- **Client deduplication minimizes network data transfers:** Data is deduplicated at the client and sent across the network in deduplicated form. In this way, only the unique data is sent to the media server, rather than the entire backup stream. Most environments – be it a LAN or a WAN environment – can benefit from less data being sent across the network. This is especially useful for WAN or Remote Office protection. Often, the data sent over the wire is a fraction of the original data size, reducing network traffic and contention for network resources.

Each Backup Exec Agent for Windows Systems has the built-in capability to do client deduplication, and many of the Application Agents (Backup Exec Agent for Exchange, Backup Exec Agent for SQL Server, etc.) can also utilize client deduplication for backup. Note that all deduplication operations require the Deduplication Option to be licensed on the Media Server. See Table 1 for details.



Backup Exec Agent	Client Deduplication
Windows File System Backups	✓
VMWare, Hyper-V	✓*
Exchange, SharePoint, SQL Server, Active Directory Agents	✓
Lotus Notes, Oracle on Windows, etc.	✓
Remote Windows Systems	✓
Remote Non-Windows Agents	No

Table 1 – Client Deduplication

* Note – for Client deduplication with Backup Exec Agent for VMWare (AVVI) and Hyper-V, the Backup Exec Remote Agent must be installed in the Guest Operating System. In this circumstance, the Remote Agent is responsible for performing the backup, and not the Agent for VMWare or the Agent for Hyper-V. Symantec recommends that for optimal deduplication, a Remote Agent also be installed in Hyper-V guests.

Media Server Deduplication

Media Server deduplication can be a useful and effective deduplication method for Linux and Solaris environments. This method of deduplication is entirely performed on the PowerVault DL Backup to Disk Appliance and does not impact source systems any more than a typical backup.

Media Server deduplication is optimal for situations where:

- **The remote system's processor is fully utilized:** If the remote system has no processor cycles to spare for deduplication calculations, Media Server deduplication can take the load and still perform deduplication.
- **The remote system is NetWare, Linux, or Solaris:** The Remote Agents for non-Windows platforms do not have the ability to do client deduplication. These systems can only take advantage of Media Server deduplication to greatly reduce the amount of storage necessary for backups.

Media Server deduplication is NOT recommended for the following environments:

- **Remote Office Protection over a WAN:** With Media Server Deduplication, the PowerVault DL Backup to Disk Appliance receives the entire data set before deduplication takes place. This is not a WAN-friendly method of deduplication. Generally, Remote Office protection without local storage should use Client Deduplication.

Any PowerVault DL Backup to Disk Appliance that has the Deduplication Option licensed can utilize Media Server deduplication. Most Agents and backup types supported by Backup Exec can take advantage of the space savings inherent with Media Server Deduplication.

Backup Exec Agent	Media Server Deduplication
Windows File System Backups	✓
VMWare, Hyper-V*	✓*
Exchange, SharePoint, SQL Server, Active Directory Agents	✓
Notes, Oracle on Windows, etc.	✓
Remote Windows Systems	✓
Remote Non-Windows Agents	✓

Table 2 – Media Server Deduplication

* Hyper-V backups will experience higher levels of deduplication when a Remote Agent is installed in the virtual Guest operating system. In this case, the backup is done via the Remote Agent installed in the Virtual Guest. The Agent for Hyper-V is not involved in Client deduplication.

Backup Exec's Deduplication Storage Folder

The PowerVault DL Backup to Disk Appliance automatically configures the Deduplication Storage Folder. A Deduplication Storage Folder is where all deduplication segments are stored, regardless of whether Client or Media Server deduplication was used for a specific backup. A PowerVault DL Backup to Disk Appliance will support a single Deduplication Storage Folder with a maximum capacity of 16 TB of deduplicated data.

The Deduplication Storage Folder contains two distinct items. The first is a file storage location, where the physical deduplication segments are stored. The second is a Postgres SQL database, where the deduplication segments are tracked and maintained. By default, the file storage location and the Postgres SQL database are installed to the same location. Right click on the Deduplication Storage Folder in the Backup Exec Device tab and select properties to access the default settings. The default settings specify 2 concurrent operations for the Deduplication Storage Folder. The Concurrent operations setting represent the number of backup or restore operations that the Deduplication Storage Folder will process simultaneously. In addition, the data stream chunk size found under the Advanced tab for the Deduplication Storage Folder properties can be changed from the default value of 64k. The data stream chunk size represents the size of each data chunk that Backup Exec writes to disk. While many customers will be able to use these defaults, some customers may need to change them

Deduplication Database Sizing

Generally, the Deduplication Database is a fraction of the total file storage location size. In Symantec's testing, the Deduplication Database increases linearly with total stored deduplicated data. Plan for roughly 6-8 GB of database size per 1 TB of stored deduplicated data; e.g. 8 TB of deduplicated data would equate to a 50 GB deduplication database.

Due to periodic weekly database maintenance routines, the PowerVault DL Backup to Disk Appliance requires double the database size available on disk. This is because automated database maintenance routines involve making a backup copy of the database. In the example above, where the deduplication database is 50 GB, the virtual disk holding the deduplication database needs to be at least 100 GB in size to account for maintenance activities alongside normal operation. The low disk space threshold for the Deduplication Storage Folder can be used to reserve a minimum amount of space for the database maintenance operations. The low space threshold can be modified by right clicking on the Deduplication Storage Folder from the Devices tab in Backup Exec. Select Properties from the pop-up. Select Advanced. The low disk space threshold can be modified from this menu. For data that gets manually removed, space reclamation is automatically queued for processing twice a day.

Processor Utilization with Client and Media Server Deduplication

Depending on the type of deduplication used, processor utilization will vary. In general, the deduplication process is not gated or throttled in any way, and is designed to accomplish deduplicated backups and restores as quickly as possible.

Client Deduplication performs the bulk of the deduplication calculations on the client (or source) system. The client deduplication process will consume up to one (1) core of one processor on the client system. The actual amount of processor utilization will depend on the amount of data to be deduplicated and the speed of the processor. Expect to see at least 75% utilization of the processor core for the duration of the backup.

Media Server deduplication performs the bulk of the deduplication calculations on the PowerVault DL Backup to Disk Appliance. Similar to client deduplication, the Media Server deduplication process will consume up to one (1) core of one processor on the PowerVault DL Backup to Disk Appliance. The actual amount of processor utilization will depend on the amount of data to be deduplicated and the speed of the processor. Expect to see at least 75% utilization of the processor core for the duration of any Media Server deduplication backup job. For both Client and Media Server deduplication, initial backup jobs will be the slowest. Backup speeds will increase over time as more database fingerprints are created.

For Remote Agents that cannot use Client Deduplication (Remote Agent for Linux and Unix [RALUS], Remote Agent for Mac Server [RAMS], Remote Agent for NetWare [RANW], etc.) there is no change to the system requirements as outlined in the Backup Exec 2010 Administration Guide. This also holds true for Windows Remote Agents that do not choose to use Client Deduplication.

Memory Utilization with Client and Media Server Deduplication

With both Client and Media Server deduplication, the majority of memory consumption takes place on the Media Server. This design is primarily for performance optimization, allowing for fast and accurate calculation of deduplication fingerprints. On the Media Server, both client and Media Server deduplication requires 1 GB (gigabyte) of physical memory for every 1 TB (terabyte) of deduplicated data stored by the Media Server. For example, if 8 TB of deduplicated data is stored, the Media Server would require at least 8 GB of physical memory.

Memory requirements for clients using Client deduplication are not stringent. Symantec requires 1 GB of physical memory on each individual client that uses Client Deduplication.

Remote System Requirements for Deduplication

Remote Agent Requirements

For Windows Remote Agents configured to use Client Deduplication the following minimum system requirements must be met:

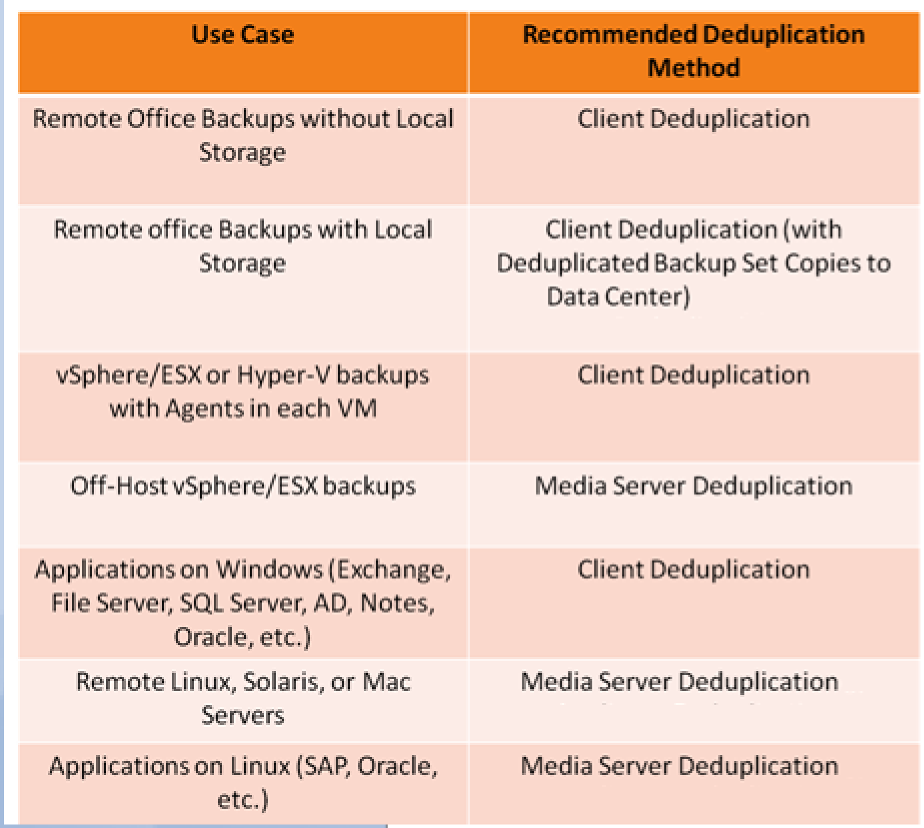
- **Processor:** For Client Deduplication, a Windows Remote Agent must have at least one dual-core processor
- **Physical Memory:** For Client Deduplication, a Windows Remote Agent must have at least 1 GB of physical system memory.

Note that Remote Agents can be **either** 64-bit or 32-bit versions of the platforms that Backup Exec supports. The Media Server has other detailed requirements listed in the Backup Exec 2010 Administration Guide. Be sure to refer to the Administration Guide before configuring Deduplication Storage Folder.

Note: When adding Remote Agents With Direct Access to the Media Server, or when the Remote Agent service on the client stops or is restarted for whatever reason, all of the BE Services on the Media Server must be restarted in order to enable direct access to the Deduplication Storage Folder.

Which Type of Deduplication to Use?

After a thorough walk-through of various deduplication methods available in Backup Exec's Deduplication Option, Administrators should have a good idea of how each type work in their environments. The table below breaks out possible use cases and environments and Symantec's recommendations for deduplication methodologies. Symantec recommends using deduplication as follows:



Use Case	Recommended Deduplication Method
Remote Office Backups without Local Storage	Client Deduplication
Remote office Backups with Local Storage	Client Deduplication (with Deduplicated Backup Set Copies to Data Center)
vSphere/ESX or Hyper-V backups with Agents in each VM	Client Deduplication
Off-Host vSphere/ESX backups	Media Server Deduplication
Applications on Windows (Exchange, File Server, SQL Server, AD, Notes, Oracle, etc.)	Client Deduplication
Remote Linux, Solaris, or Mac Servers	Media Server Deduplication
Applications on Linux (SAP, Oracle, etc.)	Media Server Deduplication

Table 3: Deduplication Methodology by Use Case

Migrating Data to Tape for Long-Term Storage

An environment with a disk-to-disk-to-tape architecture is fairly common among customers who are interested in deduplication. It's important to note that all of the forms of deduplication mentioned here are disk-based. Deduplicated data is never stored on tape in its deduplicated form; however, the process of migrating deduplicated data to tape is a simple process. Customers can set up "Set Copy" jobs that copy data from the deduplicated storage folder to another device, like a tape device, for long-term storage and retention.

These Set Copy jobs are separate jobs, but aside from the typical job configuration information needed to create a job, Backup Exec handles the details of copying deduplicated data to a tape location.
For data that was backed up using Client or Media Server deduplication, the PowerVault DL Backup to Disk Appliance will be responsible for recreating whole files from deduplicated data before transferring to tape, so there will be some impact to processor and memory usage during the Set Copy operation. While resource consumption varies based on data set, at most the Set Copy process will use 100% of one processor core while performing the Set Copy.

Restoring Deduplicated Data

Restoring deduplicated data follows the same process as setting up a regular restore job of backup data that has not been deduplicated. See Chapter 14 of the Backup Exec Administrator's Guide for specific details on setting up restore jobs.

Conclusion

The PowerVault DL Backup to Disk Appliance powered by Symantec Backup Exec 2010 introduces new capabilities around storage management through Deduplication. The Deduplication Option provides customers with the ability to reduce backup storage by 90% or more, improve backup windows, and facilitate better remote office protection. The Backup Exec Deduplication Option gives administrators the flexibility to choose what type of deduplication to use – whether it is software-based deduplication using Client or Media Server deduplication

Symantec Corporation World Headquarters
20330 Stevens Creek Blvd., Cupertino, CA 95014 USA
+1 (408) 517 8000 / +1 (800) 721 3934
www.symantec.com