HP Data Protector

Backup Strategy Guide

Technical white paper

Table of contents

Abstract

This technical white paper details backup strategies for HP Data Protector and offers an in-depth technical discussion of its development and performance expectations. In general, the white paper's purpose is to provide an organization's technologists with a technical overview and to discuss the architecture of the product or solution, including HP's technology approach.

Introduction

Backup strategies are essential for today's businesses. Backed up data serves two distinct purposes:

1. The primary purpose is to recover data as a reaction to data loss, whether caused by data deletion or corrupted data. Data loss is a very common experience of computer users and businesses. The U.S. Bureau of Labor reported that 93 percent of companies that suffer a significant data loss are out of business within 5 years.
2. The secondary purpose of backups is to recover data from a historical period of time within the constraints of a user-defined data retention policy, typically configured within a backup application for how long copies of data are required.

Though backups popularly represent a simple form of disaster recovery and should be part of a disaster recovery plan, by themselves, backups should not alone be considered disaster recovery. The U.S. National Fire Protection Agency reported that 43% of businesses that experience a disaster or emergency never reopen and a minimum of 29% close within three years.

Since a backup system contains at least one copy of all data worth saving, the data storage requirements are considerable. Organizing this storage space and managing the backup process is a complicated undertaking. A data repository model can be used to provide structure to the storage. In the modern era of computing there are many different types of data storage devices that are useful for making backups. There are also many different ways in which these devices can be arranged to provide geographic redundancy, data security, and portability.

Before data is sent to its storage location, it is selected, extracted, and manipulated. Many different techniques have been developed to optimize the backup procedure. These include optimizations for dealing with open files and live data sources, as well as compression, encryption, and de-duplication, among others. Many organizations and individuals try to have confidence that the process is working as expected and work to define measurements and validation techniques. It is also important to recognize the limitations and human factors involved in any backup scheme.

Storage concepts

Data repository models

Any backup strategy starts with a concept of a data repository. The backup data needs to be stored somehow, and probably should be organized to a degree. It can be as simple as a sheet of paper with a list of all backup tapes and the dates they were written, or a more sophisticated setup with a computerized index, catalog, or relational database. Different repository models have different advantages. This is closely related to choosing a backup rotation scheme.

Unstructured

An unstructured repository may simply be a stack of USB or CD/DVD media with minimal information about what was backed up and when. This is the easiest to implement, but probably the least likely to achieve a high level of recoverability.

Full only/system imaging

A repository of this type contains complete system images from one or more specific points in time. This technology is frequently used by computer technicians to record known good configurations. Imaging is generally more useful for deploying a standard configuration for many systems rather than as a tool for making ongoing backups of diverse systems.

Incremental/differential

An incremental style repository aims to make it more feasible to store backups from more points in time by organizing the data into increments of change between points in time. This eliminates the need to store duplicate

copies of unchanged data. Typically, at the start, a full backup (of all files) is made. After that, any number of incremental or differential backups can be made. Restoring a whole system to a certain point in time requires locating the last full backup taken previous to that time, and all the incremental/differential backups that cover the period of time between the full backup and the particular point in time to which the system is supposed to be restored. Additionally, some backup systems, such as Data Protector, can reorganize the repository to synthesize full backups from a series of incrementals.

Note: Different implementations of backup systems frequently use specialized or conflicting definitions of these terms. The most relevant characteristic of a type of incremental backup is which reference point is used to check for changes. By a common definition, a differential backup copies files that have been created or changed since the last full backup, regardless of whether any other backups have been since then, and an incremental backup refers to a backup that only includes changes made since the most recent backup of any type. Other variations include multi-level incrementals and incremental backups that compare parts of files instead of just the whole file.

Continuous data protection

Instead of scheduling periodic backups, the system immediately logs every change on the host system. This is generally done by saving byte or block-level differences rather than file-level differences. It differs from simple disk mirroring in that it enables a roll-back of the log and thus restoration of an old image of data.

Storing your backups

Regardless of the repository model that is used, the data has to be stored on some data storage medium somewhere.

Magnetic tape

Magnetic tape has long been the most commonly used medium for bulk data storage, backup, archiving, interchange, and off-site storage. Tape has typically had a capacity/price ratio that is an order of magnitude better than hard disk, but recently the ratios for tape and hard disk have become a lot closer. There are myriad formats, many of which are proprietary or specific to certain markets like mainframes or a particular brand of personal computer. Tape is a sequential access medium, so even though access times may be poor, the rate of continuously writing or reading data can actually be very fast. Some new tape drives with built-in compression and encryption are even faster than modern hard disks. A principal advantage of tape is that it has been used for this purpose for decades (much longer than any alternative) and its characteristics are well understood.

Hard disk drive

The capacity/price ratio of hard disk drives (HDDs) has been rapidly improving for many years. This is making it more competitive when compared with magnetic tape as a bulk storage medium. The main advantages of hard disk storage are low access times, availability, capacity and ease of use. External disks can be connected via local interfaces like SCSI, USB, FireWire, or eSATA, or via longer distance technologies like Ethernet, iSCSI, or Fibre Channel. Some disk-based backup systems, such as Virtual Tape Libraries, support data deduplication which can dramatically reduce the amount of disk storage capacity consumed by daily and weekly backup data. The main disadvantages of hard disk backups are that they are easily damaged, especially while being transported (for instance, for off-site backups), and that their stability over periods of years is a relative unknown.

Solid state storage

Also known as flash memory, thumb drives, USB flash drives, CompactFlash, SmartMedia, Memory Stick, Secure Digital cards, and so on, these devices are relatively costly for their low capacity, but offer excellent portability and ease-of-use.

Solid state drive

A solid-state drive (SSD) is a data storage device that uses solid-state memory to store persistent data with the intention of providing access in the same manner of a traditional block I/O hard disk drive. SSDs are different from traditional hard disk drives (HDDs), which are electromechanical devices containing spinning disks and movable read/write heads. SSDs, in contrast, use microchips that retain data in non-volatile memory chips and contain no moving parts. Compared to electromechanical HDDs, SSDs are typically less susceptible to physical shock, are silent, and have lower access time and latency, but are more expensive per gigabyte and typically support a limited number of writes over the life of the device. SSDs use the same interface as hard disk drives, thus easily replacing them in most applications.

Hybrid drive

A hybrid drive combines the features of an HDD and an SSD in one unit, containing a large HDD, with a smaller SSD cache to improve performance of frequently accessed files. These can offer near-SSD performance in most

applications (such as system startup and loading applications) at a lower price than an SSD. These are not suitable for data-intensive work, nor do they offer the other advantages of SSDs.

Optical storage

Recordable CDs, DVDs, and Blu-ray Discs are commonly used with personal computers and generally have low media unit costs. However, the capacities and speeds of these and other optical disks are typically an order of magnitude lower than hard disk or tape. Many optical disk formats are of the WORM (write once, read many) type, which makes them useful for archival purposes since the data cannot be changed. The use of an auto-changer or jukebox can make optical disks a feasible option for larger-scale backup systems. Some optical storage systems allow for cataloged data backups without human contact with the disks, allowing for longer data integrity.

Remote backup service and cloud storage

As broadband internet access becomes more widespread, remote backup services are gaining in popularity. Backing up via the internet to a remote location can protect against some worst-case scenarios such as fires, floods, or earthquakes which would destroy any backups in the immediate vicinity along with everything else. There are, however, a number of drawbacks to remote backup services. First, internet connections are usually slower than local data storage devices. Residential broadband is especially problematic as routine backups must use an upstream link that is usually much slower than the downstream link used only occasionally to retrieve a file from backup. This tends to limit the use of such services to relatively small amounts of high value data. Secondly, users must trust a third-party service provider to maintain the privacy and integrity of their data, although confidentiality can be assured by encrypting the data before transmission to the backup service with an encryption key known only to the user. Ultimately the backup service must itself use one of the above methods, so this could be seen as a more complex way of doing traditional backups.

Managing the data repository

Regardless of the data repository model or data storage media used for backups, a balance needs to be struck between accessibility, security and cost. These media management methods are not mutually exclusive and are frequently combined to meet the needs of the situation. Using on-line disks for staging data before it is sent to a near-line tape library is a common example.

Online

Online backup storage is typically the most accessible type of data storage, which can begin to restore data in milliseconds. A good example would be an internal hard disk or a disk array (maybe connected to SAN). This type of storage is very convenient and speedy, but is relatively expensive. Online storage is quite vulnerable to being deleted or overwritten, either by accident, by intentional malevolent action, or in the wake of a data-deleting virus payload.

Nearline

Nearline storage is typically less accessible and less expensive than online storage, but still useful for backup data storage. A good example would be a tape library with restore times ranging from seconds to a few minutes. A mechanical device is usually involved in moving media units from storage into a drive where the data can be read or written. Generally it has safety properties similar to online storage.

Offline

Offline storage requires some direct human action in order to make access to the storage media physically possible. This action is typically inserting a tape into a tape drive or plugging in a cable that allows a device to be accessed. Because the data is not accessible via any computer except during limited periods in which it is written or read back, it is largely immune to a whole class of online backup failure modes. Access time will vary depending on whether the media is on-site or off-site.

Off-site data protection

In computing, off-site data protection, or vaulting, is the strategy of sending critical data out of the main location (off the main site) as part of a disaster recovery plan. Data is usually transported off-site using removable storage media such as magnetic tape or optical storage. Data can also be sent electronically via a remote backup service, which is known as electronic vaulting or e-vaulting. Sending backups off-site ensures that systems and servers can be reloaded with the latest data in the event of a natural disaster, accidental error, or system crash. Sending backups off-site also ensures that there is a copy of pertinent data that is not stored on-site. Off-site backup services are convenient for companies that backup pertinent data on a daily basis (classified and unclassified).

Backup site or disaster recovery center (DR center)

In the event of a disaster, the data on backup media will not be sufficient to recover. Computer systems onto which the data can be restored and properly configured networks are necessary too. Some organizations have their own data recovery centers that are equipped for this scenario. Other organizations contract this out to a third-party recovery center. Because a DR site is itself a huge investment, backing up is very rarely considered the preferred method of moving data to a DR site. A more typical way would be remote disk mirroring, which keeps the DR data as up to date as possible.

Selection and extraction of data

A successful backup job starts with selecting and extracting coherent units of data. Most data on modern computer systems is stored in discrete units, known as files. These files are organized into filesystems. Files that are actively being updated can be thought of as "live" and present a challenge to back up. For disaster recovery, it is also useful to save metadata that describes the computer or the filesystem being backed up.

Deciding what to back up at any given time is a harder process than it seems. By backing up too much redundant data, the data repository will fill up too quickly. Backing up an insufficient amount of data can eventually lead to the loss of critical information.

Files

Making copies of files is the simplest and most common way to perform a backup. A means to perform this basic function is included in all backup software and all operating systems.

Filesystems

Filesystem dump

Instead of copying files within a filesystem, a copy of the whole filesystem itself can be made. This is also known as a raw partition backup and is related to disk imaging. The process usually involves unmounting the filesystem and running a program like `dd` (UNIX). Because the disk is read sequentially and with large buffers, this type of backup can be much faster than reading every file normally, especially when the filesystem contains many small files, is highly fragmented, or is nearly full. But because this method also reads the free disk blocks that contain no useful data, this method can also be slower than conventional reading, especially when the filesystem is nearly empty.

Identification of changes

Some filesystems have an archive bit for each file that says it was recently changed. The backup software looks at the date of the file and compares it with the last backup to determine whether the file was changed.

Application and database data

If a computer system is in use while it is being backed up, the possibility of files being open for reading or writing is real. If a file is open, the contents on disk may not correctly represent what the owner of the file intends. This is especially true for database files of all kinds. The term fuzzy backup can be used to describe a backup of live data that looks like it ran correctly, but does not represent the state of the data at any single point in time. This is because the data being backed up changed in the period of time between when the backup started and when it finished. For databases in particular, fuzzy backups are worthless.

Snapshot backup

A snapshot is an instantaneous function of some storage systems that presents a copy of the file system as if it were frozen at a specific point in time, often by a copy-on-write mechanism. An effective way to back up live data is to temporarily quiesce it (stopping file I/O), take a snapshot, and then resume live operations. At this point the snapshot can be backed up through normal methods. While a snapshot is very handy for viewing a filesystem as it was at a different point in time, it is hardly an effective backup mechanism by itself.

Open file backup

Many backup software packages feature the ability to handle open files in backup operations. Some simply check for openness and try again later. File locking is useful for regulating access to open files.

When attempting to understand the logistics of backing up open files, one must consider that the backup process could take several minutes to back up a large file such as a database. In order to back up a file that is in use, it is vital that the entire backup represents a single-moment snapshot of the file, rather than a simple copy of a read-

through. This represents a challenge when backing up a file that is constantly changing. Either the database file must be locked to prevent changes, or a method must be implemented to ensure that the original snapshot is preserved long enough to be copied, all while changes are being preserved. Backing up a file while it is being changed (in a manner that causes the first part of the backup to represent data *before* changes occur to be combined with later parts of the backup *after* the change) results in a corrupted file that is unusable as most large files contain internal references between their various parts that must remain consistent throughout the file.

Cold database backup

During a cold backup, the database is closed or locked and not available to users. The data files do not change during the backup process so the database is in a consistent state when it is returned to normal operation.

Hot database backup

Some database management systems offer a means to generate a backup image of the database while it is online and usable ("hot"). This usually includes an inconsistent image of the data files plus a log of changes made while the procedure is running. Upon a restore, the changes in the log files are reapplied to bring the database in sync.

Metadata

Not all information stored on the computer is stored in files. Accurately recovering a complete system from scratch requires keeping track of this non-file data too.

Boot sector

The boot sector can sometimes be recreated more easily than saving it. However, it usually is not a normal file and the system will not boot without it.

Partition layout

The layout of the original disk, as well as partition tables and filesystem settings, is needed to properly recreate the original system.

File metadata

Each file's permissions, owner, group, ACLs, and any other metadata need to be backed up for a restore to properly recreate the original environment.

System metadata

Different operating systems have different ways of storing configuration information. Microsoft Windows keeps a registry of system information that is more difficult to restore than a typical file.

Data manipulation and optimization

It is frequently useful or necessary to manipulate the data being backed up to optimize the backup process. These manipulations can provide many benefits including improved backup speed, restore speed, data security, media usage and/or reduced bandwidth requirements.

Compression

Various schemes can be employed to shrink the size of the source data to be stored so that it uses less storage space. Compression is frequently a built-in feature of tape drive hardware.

Deduplication

When multiple similar systems are backed up to the same destination storage device, there exists the potential for much redundancy within the backed-up data. For example, if 20 Windows workstations are backed up to the same data repository, they may share a common set of system files. The data repository only needs to store one copy of those files to be able to restore any one of the workstations. This technique can be applied at the file level or even on raw blocks of data, potentially resulting in a massive reduction in the storage space required. Deduplication can occur on a server before any data moves to backup media, sometimes referred to as source/client side deduplication. This approach also reduces the bandwidth required to send backup data to its target media. The process can also occur at the target storage device, sometimes referred to as inline or back-end deduplication.

Duplication

Sometimes backup jobs are duplicated to a second set of storage media. This can be done to rearrange the backup images to optimize restore speed, or to have a second copy at a different location or on a different storage medium.

Encryption

High-capacity removable storage media such as backup tapes present a data security risk if they are lost or stolen. Encrypting the data on these media can mitigate this problem, but presents new problems. Encryption is a CPU intensive process that can slow down backup speeds, and the security of the encrypted backups is only as effective as the security of the key management policy.

Multiplexing

When there are many more computers to be backed up than there are destination storage devices, the ability to use a single storage device with several simultaneous backups can be useful.

Staging

Sometimes backup jobs are copied to a staging disk before being copied to tape. This process is sometimes referred to as D2D2T, an acronym for Disk to Disk to Tape. This can be useful if there is a problem matching the speed of the final destination device with the source device as is frequently faced in network-based backup systems. It can also serve as a centralized location for applying other data manipulation techniques.

Managing the backup process

It is important to understand that backing up is a process. As long as new data is being created and changes are being made, backups will need to be updated. Individuals and organizations with anything from one computer to thousands (or even millions) of computer systems all have requirements for protecting data. While the scale is different, the objectives and limitations are essentially the same. Likewise, those who perform backups need to know to what extent they were successful, regardless of scale.

Objectives

Recovery point objective (RPO)

The RPO is the point in time that the restarted infrastructure will reflect. Essentially, this is the roll-back that will be experienced as a result of the recovery. The most desirable RPO would be the point just prior to the data loss event. Making a more recent recovery point achievable requires increasing the frequency of synchronization between the source data and the backup repository.

Recovery time objective (RTO)

The RTO is the amount of time that elapses between the disaster and restoration of business functions.

Data security

In addition to preserving access to data for its owners, data must be restricted from unauthorized access. Backups must be performed in a manner that does not compromise the original owner's undertaking. This can be achieved with data encryption and proper media handling policies.

Limitations

An effective backup scheme will take into consideration the limitations of the situation.

Backup window

The period of time when backups are permitted to run on a system is called the backup window. This is typically the time when the system sees the least usage and the backup process will cause the least amount of interference with normal operations. The backup window is usually planned with users' convenience in mind. If a backup extends past the defined backup window, a decision must be made whether it is more beneficial to abort the backup or to lengthen the backup window.

Performance impact

All backup schemes have some performance impact on the system being backed up. For example, for the period of time that a computer system is being backed up, the hard drive is busy reading files for the purpose of backing up, and its full bandwidth is no longer available for other tasks. Such impacts should be analyzed.

Costs of hardware, software, and labor

All types of storage media have a finite capacity with a real cost. Matching the correct amount of storage capacity (over time) with the backup needs is an important part of the design of a backup scheme. Any backup scheme has some labor requirement, but complicated schemes have considerably higher labor requirements. The cost of backup software must also be considered.

Network bandwidth

Distributed backup systems can be affected by limited network bandwidth (SAN/LAN/WAN).

Implementation

Meeting the defined objectives in the face of the above limitations can be a difficult task. The following tools and concepts can make that task more achievable.

Scheduling

Using a job scheduler can greatly improve the reliability and consistency of backups by removing part of the human element.

Authentication

Over the course of regular operations, the user accounts and/or system agents that perform the backups need to be authenticated at some level. The power to copy all data off or onto a system requires unrestricted access. Using an authentication mechanism is a good way of preventing the backup scheme from being used for unauthorized activity.

Chain of trust

Removable storage media are physical items and must only be handled by trusted individuals. Establishing a chain of trusted individuals (and vendors) is critical to defining the security of the data.

Measuring the process

To ensure that the backup scheme is working as expected, the process needs to include the monitoring of key factors and maintenance of historical data.

Backup validation

This is the process by which owners of data can get information about how their data was backed up. It is also used to prove compliance to regulatory bodies outside the organization. For example, under the US Health Insurance Portability and Accountability Act (HIPAA), an insurance company might be required to show ”proof“ that their patient data is meeting records retention requirements. Disaster, data complexity, data value and increasing dependence upon ever-growing volumes of data all contribute to the anxiety around and dependence upon successful backups to ensure business continuity. For that reason, many organizations rely on third-party or ”independent“ solutions to test, validate, and optimize their backup operations (backup reporting).

Reporting

In larger configurations, reports are useful for monitoring media usage, device status, errors, vault coordination and other information about the backup process.

Logging

In addition to the history of computer-generated reports, activity and change logs are useful for monitoring backup system events.

Data validation

Backup programs make use of checksums or hashes to validate that the data was accurately copied. They allow data integrity to be verified without reference to the original file; if the file as stored on the backup medium has the same checksum as the saved value, then it is very probably correct.

Monitored backup

Backup processes are monitored by a third-party monitoring center. This center alerts users to any errors that occur during automated backups. Monitoring services also allow the collection of historical metadata, which can be used for Storage Resource Management (SRM) purposes such as the projection of data growth, locating redundant primary storage capacity and determining reclaimable backup capacity.

Disaster recovery

Disaster recovery is the process, policies and procedures related to preparing for recovery or continuation of the technology infrastructure critical to an organization after a natural or human-induced disaster. Disaster recovery is a subset of business continuity. While business continuity involves planning for keeping all aspects of a business functioning in the midst of disruptive events, disaster recovery focuses on the IT or technology systems that support business functions.

Importance of planning

As IT systems have become increasingly critical to the smooth operation of a company, and arguably the economy as a whole, the importance of ensuring the continued operation of those systems, and the rapid recovery of the systems, has increased. Backups are the foundation of recoveries and better backups can be achieved through better planning.

Strategies

Before selecting a disaster recovery strategy, a disaster recovery planner should refer to their organization's business continuity plan, which should indicate the key metrics of recovery point objectives (RPOs) and recovery time objectives (RTOs) for various business processes (such as the process for running payroll, generating an order, and so on). The metrics specified for the business processes must then be mapped to the underlying IT systems and infrastructure that support those processes.

Once the RTO and RPO metrics have been mapped to IT infrastructure, the disaster recovery planner can determine the most suitable recovery strategy for each system. An important note here however is that the business ultimately sets the IT budget and therefore the RTO and RPO metrics need to fit with the available budget. While most business unit heads would like zero data loss and zero time loss, the cost associated with that level of protection may make the desired high-availability solutions impractical.

The following is a list of the most common strategies for data protection:

- Backups made to tape and sent off-site at regular intervals
- Backups made to disk on-site and automatically copied to off-site disk and finally copied to tape
- Replication of data to an off-site location, which overcomes the need to restore the data (only the systems need to be restored or synchronized)
- High-availability systems that keep both the data and system replicated off-site, enabling continuous access to systems and data

Planning your backup strategy

Planning is the most important task for a successful implementation of Data Protector. It is also a prerequisite for setting up adequate backup equipment, such as the number of backup server and storage devices.

Defining the requirements

You should ask certain questions when planning a backup strategy, even if a backup concept is already available. The questions should at least reveal risks and consequences.

Please consider that the planning should not be price-based. That would typically result in committing backup throughput and storage capacity but not, for instance, the point-in-time recoverability of critical applications.

Answering the right questions is the foundation of the requirement definition. The most important questions are:

- Why does your business need backups? Are there any business policies and service level agreements (SLAs) regarding backup and restore?

  Some organizations have defined policies for storing and archiving data. Your backup strategy should comply with these policies.

- What types of data need to be backed up?

  List all types of data existing in your network, such as user files, system files, and databases. List also the operating system used and Hypervisor, if applicable.

- How much data should be backed up?

  For each server and client, list the type of data, its data volume, and the expected growth rate.

- How should the data be backed up?

  Does the business allow cold or only hot database backups? Are there any performance critical systems that must not be backed up directly? Do you consider backups from replicated data?

- When and how often do you need backups?

  For each type of data, list how often the data needs to be backed up. For example, user files may be backed up daily, system data weekly, and some database transactions twice a day.

  Note: The backup frequency will determine the maximum possible data loss. If a filesystem backup is scheduled daily, the maximum possible data loss is one day.

- How long can a backup and restore take?

  For each server and client, list the type of data and the maximum backup/restore time. If there are any relevant SLAs or policies, add the name (for example, gold, silver, and bronze).

- What is the maximum recovery time?

  For each server and client, list the type of data and the maximum recovery time. This is the time before the server and client is available again for the business.

  Note: Database recoveries with many transaction log files could take a considerable time.

- Where should the data be stored and how long should it be kept?

  List how long each type of data must be available and protected. If applicable, consider retention times for online, nearline and off-site data storage.

- Do you need media duplicates?

  List all types of data that should be protected by duplicated media. Duplicates could prevent data loss in the case of damaged media.

- What about excluding unnecessary data?

  Most businesses do not need to know all details of backed-up files (for example, the owner, modification data, permission), so the tracking could be changed to a lower level. This would reduce the size of the file catalog.

  Many businesses are facing huge space requirements for user-downloaded files, such as videos. Usually, a backup is not required because they could be downloaded again.

  Another example is temporary files and directories, which could be excluded to reduce backup volumes and runtimes.

- What about security?

  Who can administer or operate a backup application (Data Protector)?

  Who can physically access client systems and backup media?

  Who can restore data?

  Who can view information about backed-up data?

- How is the network designed?

  What about the network infrastructure?

  Is there any detailed network chart?

  Is there any SAN chart that includes all shared storage?

  Is there any separate backup LAN?

Note: It is not necessary and sometimes not possible to answer all these questions. Add all assumptions and ensure they are approved by the customer.

Documentation

All requirements should be documented and signed by affected business and IT managers. Add the documentation to any subsequent backup concept or sales proposal.

Planning for Data Protector

You need to understand your environment and how Data Protector will be used. Every component of Data Protector has particular requirements that have to be reflected in the overall solution planning.

Each Data Protector component has unique properties for you to learn, and as your knowledge of Data Protector grows, so will your expertise in implementing a successful solution.

The following subsections summarize the most important features.

Overview

HP Data Protector is a backup solution that provides reliable data protection and high accessibility for your fast growing business data. Data Protector offers comprehensive backup and restore functionality specifically tailored for enterprise-wide and distributed environments.

As shown in Figure 1, Data Protector is organized into several functional components.

Figure 1: Data Protector overview

- Scalable and highly flexible architecture
- Easy central administration
- High-performance backup
- Data security
- Support of mixed environments
- Easy installation for mixed environments
- High-availability support
- Backup object operations
- Easy restore
- Automated and unattended operation
- Service management
- Monitoring, reporting and notification
- Integration with online applications
- Integration with other products

Scalable and highly flexible architecture

Data Protector can be used in environments ranging from a single system to thousands of systems on several sites. Due to the network component concept of Data Protector, elements of the backup infrastructure can be placed in the topology according to user requirements. The numerous backup options and alternatives for setting up a backup infrastructure allow the implementation of virtually any configuration you want. Data Protector also enables the use of advanced backup concepts, such as synthetic backup and disk staging.

Easy central administration

Through its easy-to-use graphical user interface (GUI), Data Protector allows you to administer your complete backup environment from a single system. To ease operation, the GUI can be installed on various systems to allow multiple administrators to access Data Protector via their locally-installed consoles. Even multiple backup environments can be managed from a single system. The Data Protector command-line interface allows you to manage Data Protector using scripts.

High-performance backup

Data Protector enables you to perform backup to several hundred backup devices simultaneously. It supports high-end devices in very large libraries. Various backup possibilities, such as local backup, network backup, online backup, disk image backup, synthetic backup, backup with object mirroring, and built-in support for parallel data streams, allow you to tune your backups to best fit your requirements.

Data security

To enhance the security of your data, Data Protector lets you encrypt your backups so that they become protected from others. Data Protector offers two data-encryption techniques: software-based and drive-based.

Support of mixed environments

As Data Protector supports heterogeneous environments, most features are common to the UNIX and Windows platforms. The UNIX and Windows Cell Managers can control all supported client platforms (UNIX, Windows, and Novell NetWare). The Data Protector user interface can access the entire Data Protector functionality on all supported platforms.

Easy installation for mixed environments

The Installation Server concept simplifies the installation and upgrade procedures. To remotely install UNIX clients, you need an Installation Server for UNIX. To remotely install Windows clients, you need an Installation Server for Windows. The remote installation can be performed from any client with an installed Data Protector GUI. For supported platforms for the Installation Server, see the *HP Data Protector Product Announcements, Software Notes, and References*.

High-availability support

Data Protector enables you to meet the needs for continued business operations around the clock. In today's globally-distributed business environment, company-wide information resources and customer service applications must always be available. Data Protector enables you to meet high-availability needs by:

- Integrating with clusters to ensure fail-safe operation with the ability to back up virtual nodes. For a list of supported clusters, see the *HP Data Protector Product Announcements, Software Notes, and References*.
- Enabling the Data Protector Cell Manager itself to run on a cluster.
- Supporting all popular online database Application Programming Interfaces (APIs).
- Integrating with advanced high-availability solutions, such as EMC Symmetrix, HP StorageWorks P6000 EVA Disk Array family, HP StorageWorks P9000 XP Disk Array family, or HP StorageWorks P4000 SAN solutions.
- Providing various disaster recovery methods for Windows and UNIX platforms.
- Offering methods of duplicating backed-up data during and after the backup to improve fault tolerance of backups or for redundancy purposes.

Backup object operations

To provide flexibility in the choice of backup and archive strategy, advanced techniques are available for performing operations on individual backup objects. These include copying objects from one medium to another, useful for disk staging and archiving purposes, and consolidation of multiple object versions from incremental backups into a single full-backup version. To support such functionality, there is also the ability to verify both original and copied or consolidated backup objects.

Easy restore

Data Protector includes an internal database that keeps track of data, such as which files from which system are kept on a particular medium. In order to restore any part of a system, simply browse the files and directories. This provides fast and convenient access to the data to be restored.

Automated or unattended operation

With the internal database, Data Protector keeps information about each Data Protector medium and the data on it. Data Protector provides sophisticated media-management functionality. For example, it keeps track of how long a particular backup needs to remain available for restoring, and which media can be used or reused for backups. The support of very large libraries complements this, allowing for unattended operation over several days or weeks (automated media rotation). Additionally, when new disks are connected to systems, Data Protector can automatically detect or discover the disks and back them up. This eliminates the need to adjust backup configurations manually.

Service management

Data Protector is the first backup and restore management solution to support service management. The integration with Application Response Management (ARM) and Data Source Integration (DSI) enables powerful support of Service Level Management (SLM) and Service Level Agreements (SLA) concepts by providing relevant data to management and planning systems. The DSI integration provides a set of scripts and configuration files from which users are able to see how to add their own queries using Data Protector reporting capabilities.

Monitoring, reporting and notification

Superior web reporting and notification capabilities allow you to easily view the backup status, monitor active backup operations, and customize reports. Reports can be generated using the Data Protector GUI, or using the `omnirpt` command on systems running UNIX or Windows, as well as using Java-based online generated web reports. You can schedule reports to be issued at a specific time, or to be attached to a predefined set of events, such as the end of a backup session or a mount request. In addition, the Data Protector auditing functionality enables you to collect a subset of backup session information and provides an overview of backup operations. Backup session information is recorded to the audit log files.

Integration with online applications

Data Protector provides online backup of Microsoft Exchange Server, Microsoft SQL Server, Microsoft SharePoint Server, Oracle, Informix Server, SAP R/3, SAP MaxDB, Lotus Notes/Domino Server, IBM DB2 UDB, Sybase database objects, and VMware Virtual Infrastructure and Hyper-V objects. For a list of supported versions for a particular operating system, consult the Data Protector home page at http://www.hp.com/go/dataprotector.

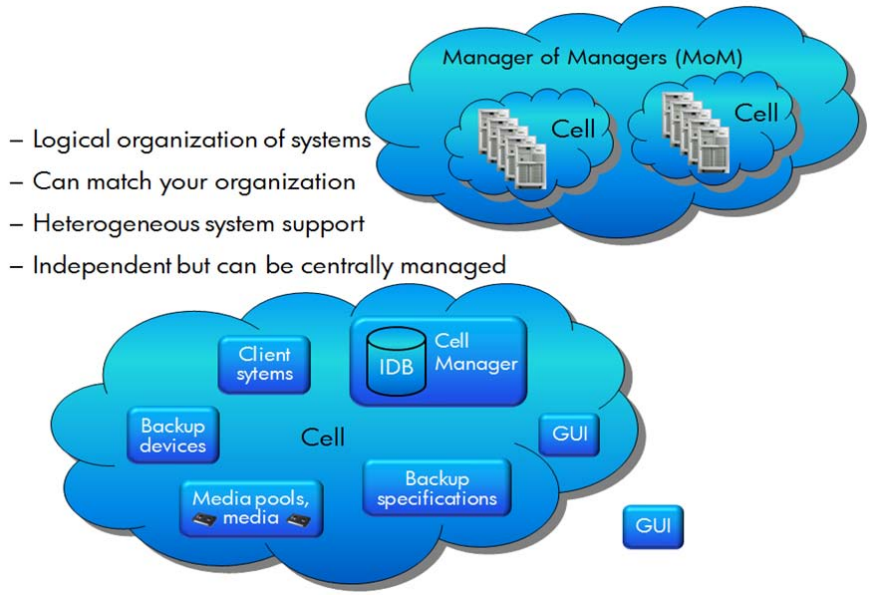
Integration with other products

Additionally, Data Protector integrates with EMC Symmetrix, Microsoft Cluster Server, MC/ServiceGuard and other products. For detailed documentation describing the features of Data Protector, including integrations, as well as the latest platform and integration support information, consult the Data Protector home page at http://www.hp.com/go/dataprotector.

Data Protector architecture

The Data Protector cell, shown in Figure 2, protects a backup domain consisting of logically organized systems. Multiple cells are managed with the Data Protector Manager-of-Managers (MoM).

Figure 2: Data Protector architecture – backup domain



Cell

The Data Protector cell, shown in Figure 3, is a backup environment that has a Cell Manager, client systems, and backup devices. A graphical user interface (GUI) manages the backup environment.

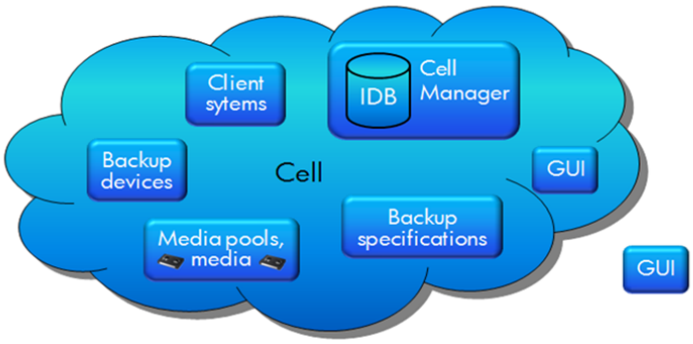
Figure 3: Data Protector architecture – cell

– Cell Manager

– Client systems, systems to be backed up

– Backup devices, connected to backup systems (tape, disk, etc.)

– GUI, graphical user interface



Cell Manager

The Cell Manager, shown in Figure 4, manages the backup environment from a central point. Backup specifications, backup devices, and media pools are all managed from the central point. The GUI could be in the same cell or outside the managed backup environment.

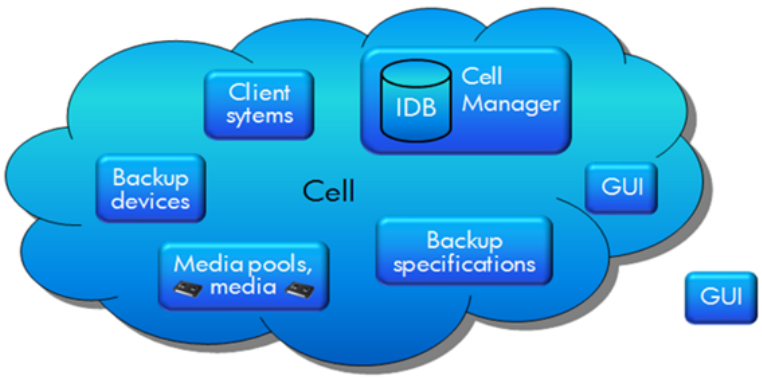
The Data Protector internal database (IDB) is part of the Cell Manager, and keeps track of the files you back up so that you can browse and easily recover the entire system or single files.

Figure 4: Data Protector architecture – Cell Manager

– Manages cell from a central point

– Configures backups, devices, media

– Contains internal database (IDB), keeps track of the files you back up



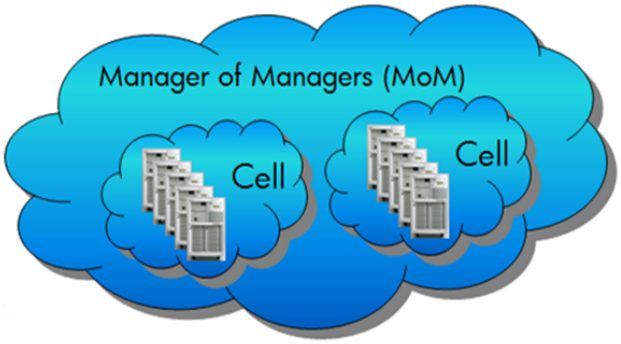
Note: The GUI and the Cell Manager systems can run on UNIX and Windows operating systems; they do not have to run the same operating system as client systems. For a list of supported operating systems for a particular Data Protector component, see http://www.hp.com/go/dataprotector.

Manager-of-Managers (MoM)

For multi-site operations, the Data Protector Manager-of-Managers (MoM) allows unlimited growth of the backup environments. MoM, shown in Figure 5, enables distribution of control to local administrators while maintaining the ability to set overall policies and monitor the entire enterprise backup environment from a central point. As a result, corporate policies can be implemented without diminishing local control and responsibility, while data protection and storage management costs do not change. Using MoM, several Data Protector cells (which are also referred to as domains) can be grouped together, configured and managed from a central cell. Using the MoM interface allows customers to scale a Data Protector backup environment to include hundreds of Disk Agents and Media Agents.

Figure 5: Data Protector architecture –Manager-of-Managers

– Groups several Data Protector cells
– Configures and manages from a central point
– For multi-site operations
– Allows unlimited growth of backup environment



Systems to be backed up

Client systems you want to back up must have the Data Protector Disk Agent (DA), also called Backup Agent, installed. To back up online database integrations, install the Application Agent. In the rest of this paper, the term Disk Agent will be used for both agents. The Disk Agent reads or writes data from a disk on the system and sends or receives data from a Media Agent. The Disk Agent is also installed on the Cell Manager, thus allowing you to back up data from the Cell Manager, including the Data Protector configuration, and the IDB.

Systems with backup devices

Client systems with connected backup devices must have a Data Protector Media Agent (MA) installed. Such client systems are also called backup or media servers. A backup device can be connected to any system and not only to the Cell Manager. A Media Agent reads or writes data from or to media in the device and sends or receives data from the Disk Agent.

Systems with a user interface

You can manage Data Protector from any system on the network on which the Data Protector graphical user interface (GUI) and command line interface (CLI) are installed. As a result, you can have the Cell Manager system in a data center while managing Data Protector from your desktop system.

Installation Server

The Installation Server holds a repository of the Data Protector software packages for a specific architecture. The Cell Manager is by default also an Installation Server. At least two Installation Servers are needed for mixed environments: one for UNIX and one for Windows.

Data Protector key features

Central Administration GUI

Through its easy-to-use graphical user interface (GUI), Data Protector allows you to administer your complete backup environment from a single system. To ease operation, the GUI can be installed on various systems to allow multiple administrators to access Data Protector via their locally-installed consoles. Even multiple DP cells can be managed from a single GUI. The Data Protector command-line interface allows you to manage Data Protector using scripts.

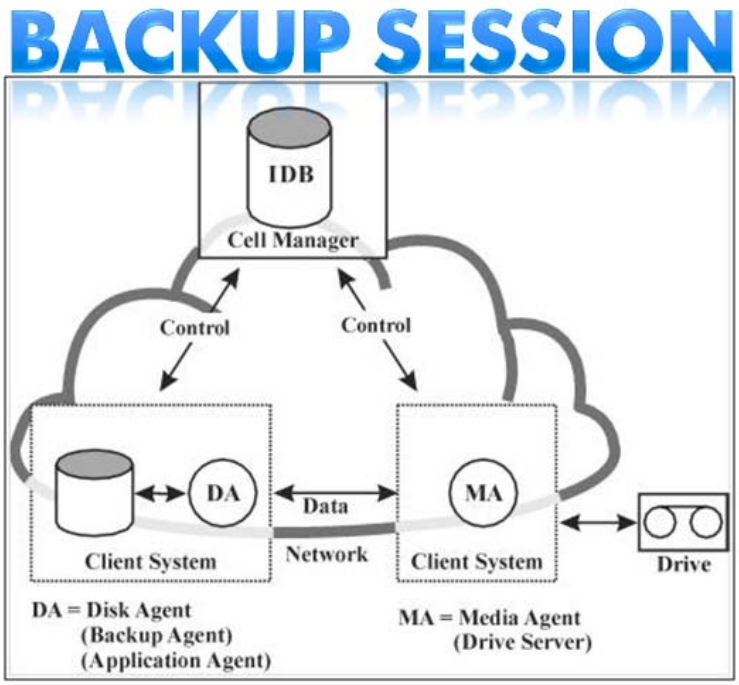
Operations in the cell

The Data Protector Cell Manager controls backup and restore sessions, which perform all the required actions for a backup or restore.

Backup session

A backup session, shown in Figure 6, is a process that creates a copy of data on storage media. It is started either interactively by an operator using the Data Protector user interface, or unattended using the Data Protector scheduler.

Figure 6: Backup session



The backup session process starts Media Agents and Disk Agents, controls the session, and stores generated messages in the IDB. Data is read by the Disk Agent and sent to a Media Agent, which saves it to media.

A typical backup session is more complex than the one shown in Figure 7. A number of Disk Agents read data from multiple disks in parallel and send data to one or more Media Agents.

Restore session

A restore session, shown in Figure 6, is a process that restores data from previous backups to a disk. The restore session is interactively started by an operator using the Data Protector user interface.

After you have selected the files to be restored from a previous backup, you invoke the actual restore. The restore session starts the necessary Media Agents and Disk Agents, controls the session, and stores messages in the IDB. Data is read by a Media Agent and sent to the Disk Agent, which writes it to disks.

Figure 7: Restore session

Internal database

The internal database (IDB), located on the Cell Manager, holds information about what data is backed up, on which media it resides, the result of backup, restore, object copy, object consolidation, object verification, and media management sessions, and which devices and libraries are configured.

The information stored in the IDB enables the following:

- Backup management, configuration and scheduling
- Monitoring, check running backup sessions
- Reporting, verifying the result of backup sessions
- Media management, allocation of media during backup, object copy, and object consolidation sessions, tracking media management operations and media attributes, grouping media in different media pools, and tracking media location in tape libraries
- Encryption/decryption management, allocation of encryption keys for encrypted backup or copy sessions, and supplying the decryption key required for the restore of encrypted backup objects
- Fast and convenient restore, browsing files and directories

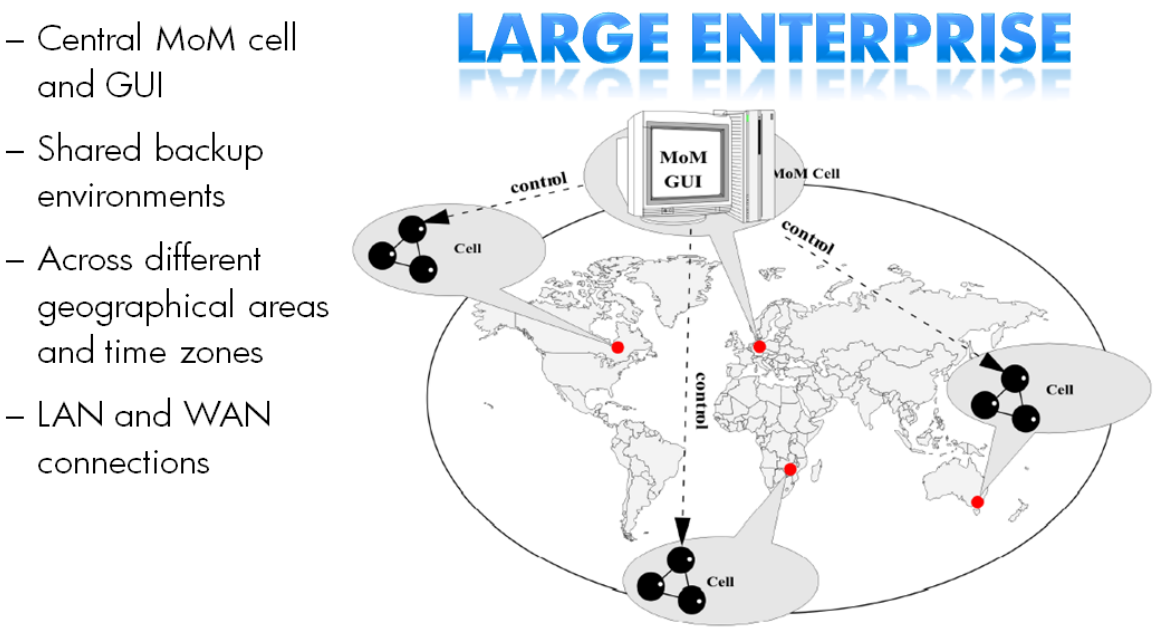
An essential part of the internal database configuration is configuring the backup of the IDB itself. Regular IDB backup is the most important preparation for recovery in the event of a disaster. The IDB recovery is essential for the restore of other backed-up data if the Cell Manager is struck by a disaster.

Enterprise environments

A typical enterprise network environment, shown in Figure 8, consists of a number of systems from different vendors with different operating systems. The systems may be located in different geographical areas and time zones. All the systems are connected with LAN or WAN networks operating at various communication speeds.

This solution can be used when several geographically separated sites require common backup policies to be used. It can also be used when all departments at the same site want to share the same set of backup devices.

Figure 8: Large Data Protector enterprise environment



The Manager-of-Managers (MoM) architecture is explained in detail below.

Manager-of-Managers (MoM) architecture

Several cells can be grouped together (joined) and managed from a central cell. The Manager-of-Managers, shown in Figure 9, provides the following features:

- Centralized licensing repository
- Simplified license management. This is optional but useful for very large environments.
- Centralized Media Management Database (CMMDB)

  The CMMDB allows you to share devices and media across several cells in a MoM environment. This makes devices of one cell (using the CMMDB) accessible to other cells that use the CMMDB. The CMMDB, if used, must reside in the MoM cell. In this case, a reliable network connection is required between the MoM cell and the other Data Protector cells. Note that it is optional to centralize the Media Management Database.

- Sharing libraries

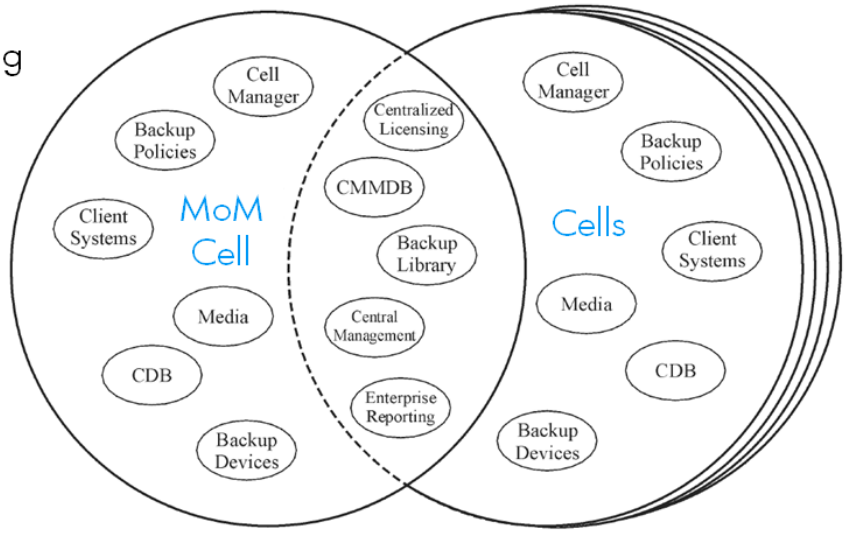
  With the CMMDB, you can share high-end devices between cells in the multi-cell environment. One cell can control the robotics, serving several devices that are connected to systems in different cells. Even the Disk Agent to Media Agent data path can go across cell boundaries.

- Enterprise reporting

  The Data Protector Manager-of-Managers can generate reports on a single-cell basis as well as for the entire enterprise environment.

Figure 9: Manager-of-Managers architecture

– Centralized licensing repository
– Centralized Media Management Database (CMMDB)
– Sharing libraries
– Enterprise reporting



A MoM environment does not require a reliable network connection from Data Protector cells to the central MoM cell, because only the controls are sent over the long-distance connection. However the backups are performed locally within each Data Protector cell. Nevertheless, this is based on the assumption that each cell has its own Media Management Database (MMDB).

Media Management

Data Protector provides you with powerful media management, which lets you easily and efficiently manage large numbers of media in your environment in the following ways:

- Grouping media into logical groups, called media pools, which allows you to think about large sets of media without having to worry about each medium individually
- Keeping track of all media and the status of each medium, data protection expiration time, availability of media for backup, and a catalog of what has been backed up to each medium
- Fully automated operation. If Data Protector controls enough media in the library devices, the media management functionality lets you run the backup sessions without operator intervention.
- Automated media rotation policies that allow media selection for backups to be performed automatically
- Recognition and support of barcodes on large library devices and silo devices with barcode support
- Recognition, tracking, viewing, and handling of media used by Data Protector in large library devices and silo devices
- The possibility of having information about the media in a central place and the sharing of this information among several Data Protector cells
- Interactive or automated creation of additional copies of the data on the media
- Support for media vaulting

Media Pool

Data Protector uses media pools to manage large numbers of media. A media pool is a logical collection of media of the same physical type with common usage policies (properties). Usage is based on the data on the media. The structure and number of pools, as well as which pool contains what type of data on its media, depend entirely on your preferences.

When a device is configured, a default media pool is specified. This media pool is used if no other media pool is defined in the backup specification.

Backup Devices

Data Protector defines and models each device as a physical device with its own usage properties, such as the default pool. This device concept is used because it allows you to easily and flexibly configure devices and use them in conjunction with backup specifications. The definition of the devices is stored in the Data Protector Media Management Database.

Figure 10: Backup specifications, backup devices and media pools

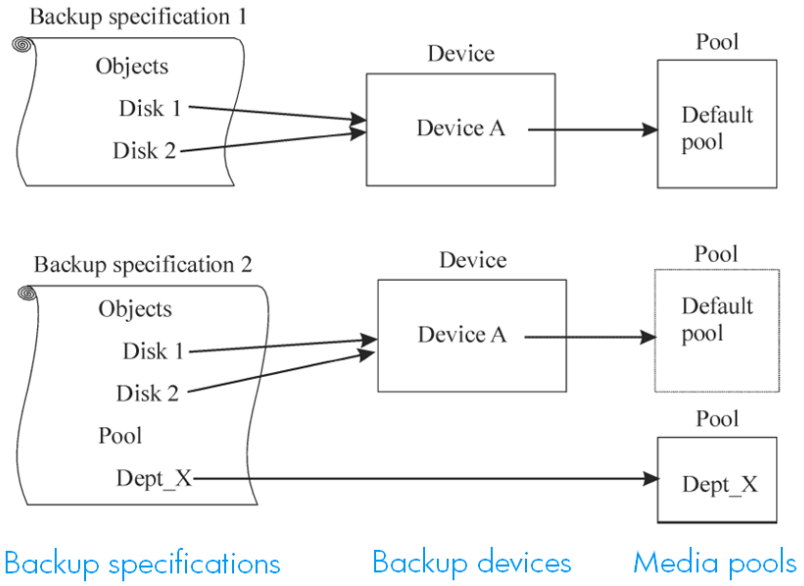


Figure 10 shows the relationship among the backup specification, devices, and media pools. The devices are referred to in the backup specification. Each device is linked to a media pool; this media pool can be changed in the backup specification. For example, *backup specification 2* references the pool *Dept_X* instead of the default pool.

Reporting

Data Protector includes robust reporting capabilities, including a notification function that allows customers to forward events to HP, or to third-party reporting or management tools such as HP Storage Essentials and IBM Tivoli.

Data Protector reports provide various details of your backup environment. For example, you can check the status of the last backup, object copy, object consolidation, or object verification, check which systems in your network are not configured for backup, check on the consumption of media in media pools, check the status of devices and more. You can configure reports and report groups using the Data Protector GUI or any Web browser with Java support. Report groups allow you to easily manage reports, to schedule the reports in the report group, and to define the criteria for grouping the reports in report groups. Parameters allow you to customize reports.

For customers who require the advanced reporting capabilities, HP provides Data Protector Reporter software. HP Data Protector Reporter software is customer-installable software which delivers centralized, automated reporting to optimize operations and infrastructure.

Data Protector Reporter features a powerful reporting engine that drives global, multi-site backup and restore analysis, and enables IT staff to reduce the risk of data loss by easily identifying and troubleshooting issues such as failed backup clients, performance issues, and drive and media utilization. Data Protector Reporter features 35+ out-of-the-box reports including SLA and performance reporting. The optional licensed module also provides enhanced powerful customized ad-hoc query and analysis flexible reporting. Collectively these reports help customer optimize operations to meet SLA, optimize capacity and performance to improve cost, plan for growth, track compliance and SLA.

Security and encryption

Security has to be planned, tested, and implemented on different security-critical layers to ensure the secure operation of Data Protector. Such layers are:

- Cell Managers

  Security of the Cell Manager can be enhanced by strict IP-checking functionality.

- Data Protector Clients

  Clients will verify the source for each request and allow only those requests received from security-enabled clients.

- Users

  Data Protector provides advanced security functionality that prevents the unauthorized backing up or restoring of data. Data Protector security involves hiding data from unauthorized users, data encoding, and restricted grouping of users according to their responsibilities.

- Data

  Data Protector lets you encrypt backed-up data so that it becomes protected from others. Two data encryption techniques are available: software-based and drive-based encryption.

  1. Data Protector software encryption, referred to as AES 256-bit encryption, is based on the AES-CTR (Advanced Encryption Standard in Counter Mode) encryption algorithm that uses random keys of 256-bit length. The same key is used for both encryption and decryption. With AES 256-bit encryption, data is encrypted before it is transferred over a network and before it is written to media.
  2. Data Protector drive-based encryption uses the encryption functionality of the drive. The actual implementation and encryption strength depend on the drive's firmware. Data Protector only turns on the feature and manages encryption keys. The key management functionality is provided by the Key Management Server (KMS), which is located on the Cell Manager. All encryption keys are stored centrally in the key store file on the Cell Manager and administered by the KMS.

- Control communication

  Data Protector encrypted control communication helps preventing unauthorized access to clients in a Data Protector cell. It is based on Secure Socket Layer (SSL), a cryptographic protocol that provides network connections and encapsulates existing Data Protector communication protocol.

Disk backup

Industry has requirements for increasingly faster methods of backing up and restoring data. In addition, it has become more and more important that the time required for data backup and restore is reduced to a minimum so as not to interrupt the day-to-day running of company applications.

Many applications and databases frequently make small changes to existing files or produce many new files containing business-critical data throughout the working day. These files need to be backed up immediately to guarantee the data in them will not be lost. This requirement means that a fast medium that can store large amounts of data that works without interruption is necessary for storing data. Disk-based storage media have become increasingly cheaper in recent years. At the same time, the storage capacity of disks has risen. This has led to the availability of low-cost, high-performance single disks and disk arrays for storing data.

Disk backup (also known as disk-to-disk backup) is becoming ever more important. In the past, tape storage was the favored medium for backup and restore because of its price and effectiveness in meeting disaster recovery requirements. Today, more and more businesses are augmenting their tape storage backup solutions with faster disk-based backup solutions. This ensures faster data backup and recovery.

Data Protector provides the following disk-based devices:

- Standalone file device

  A standalone file device is a file in a specified directory to which you back up data instead of writing to a tape. This device saves data in the form of files; each of these files is the equivalent of a slot in a tape device. The standalone file device is useful for smaller backups.

- File jukebox device

  A jukebox is a library device. It can contain either optical or file media. If the device is used to contain file media it is known as a 'file jukebox device'. The file jukebox device is a logical equivalent of a tape stack. It contains slots whose size is defined by the user during the initial device configuration. This device is configured

manually. The file jukebox properties can be altered while it is being used. If used to contain file media the device writes to disk instead of tape. The file jukebox device saves data in the form of files; each of these files is the equivalent of a slot in a tape device.

- File library device

  A file library device is a device that resides in a directory on an internal or external hard disk drive defined by you. A file library device consists of a set of directories. When a backup is made to the device, files are automatically created in these directories. The files contained in the file library directories are called file depots.

Advanced Backup to Disk

The Data Protector Advanced Backup to Disk extension provides a fast, easy way to integrate disk-based backup into your environment. Ideal if you want to stage backups on fast, centralized disk space before moving them to tape for longer-term storage, Advanced Backup to Disk enables backup to file libraries or virtual tape libraries (VTLs). Advanced Backup to Disk is perfect for remote office and branch office (ROBO) environments because it is easy to configure and maintain, and enables space optimization functionality such as synthetic and virtual full backups. In addition, together with HP StoreOnce D2D and VLS devices, Data Protector Advanced Backup to Disk provides centralized, deduplication-enabled replication management.

With Advanced Backup to Disk, you can reduce recovery times, deploy deduplication and other storage optimization techniques, improve backup performance, and build the tiered recovery architecture to suit your needs.

Device streaming and Disk Agent concurrency (multiplexing)

For maximum device performance, keep devices streaming.

A tape device streams if it can feed enough data to the medium to keep the tape moving forward continuously. Otherwise, the tape has to be stopped, the device waits for more data, reverses the tape a little, resumes writing to the tape, and so on. In other words, if the data rate written to the tape is less than or equal to the data rate that can be delivered to the device by the computer system, the device is streaming.

To allow the device to stream, a sufficient amount of data must be sent to the device. Data Protector accomplishes this by starting multiple Disk Agents for each Media Agent that writes data to the device. The number of Disk Agents started for each Media Agent is called Disk Agent (backup) concurrency. This is also known as multiplexing.

Additionally, you can concurrently back up to multiple devices. In this case, multiple Disk Agents read data in parallel and send the data to multiple Media Agents.

Zero downtime backup and instant recovery

Conventional methods of backing up to tape are not well suited to large database applications; either the database has to be taken offline or, if the application allows it, put into “hot-backup mode” while data in it is streamed to tape.

The first can cause major disruption to the application’s operation. The second can produce many large transaction log files, putting extra load on the application system.

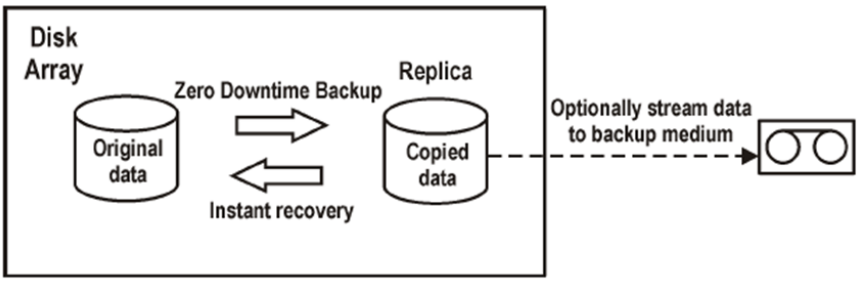
ZDB (zero downtime backup), shown in Figure 11, uses disk array technology to minimize the disruption. In very general terms, a copy or replica of the data is created or maintained on a disk array. This is very fast and has little impact on the application’s performance. The replica itself can be the backup, or it can be streamed to tape or disk without further interruption to the application’s use of the source database.

Depending on the hardware and software with which it is created, a replica may be an exact duplicate (mirror, snapclone), or a virtual copy (snapshot) of the data being backed up.

Instant recovery requires a replica to exist on the same disk array to which the data is to be restored. Application and backup systems are disabled and the contents of the replica are either restored directly to their original locations or presented to the system in place of contents of the source volumes. Because the restore is performed internally within the disk array, it runs at very high speed. Once the restore is completed, the sections of the database or filesystem concerned are returned to their states at the time the replica was created and the application system can be re-enabled.

Figure 11: Zero downtime backup and instant recovery concept

– Zero Downtime Backup (ZDB)
  • Minimal downtime or impact on the application system during backup
  • Array-based snapshot functionality
  • Backup from replica (split mirror, snapshot)
– Instant Recovery (IR)
  • Short restore times (minutes instead of hours)



Application integrations and Microsoft Volume Shadow Copy Service

In the traditional backup model, the backup application coordinates the various systems and components involved in the backup process: the application and backup systems, and the disk array.

On Windows systems, a unified backup and restore service, the Volume Shadow Copy Service (VSS) coordinates the components involved in the backup process. The VSS model, shown in Figure 12, provides a standardized interface to the applications (writers) and disk arrays (providers).

The writers interact with the applications, providing a list of items that can be backed up. Data integrity is provided by the writers on the operating system and application levels.

The hardware providers replace the disk array agent functionality and behave similarly to disk array agents from the Data Protector point of view.

When performing an instant recovery of data backed up in a zero downtime backup session with the Data Protector Microsoft Volume Shadow Copy Service integration, you can choose to use the Microsoft Virtual Disk Service (VDS) or the disk array agent. The selection also depends on the way the backup was made.

Figure 12: Data Protector application integration with Microsoft VSS

Disaster recovery

A computer disaster refers to any event that renders a computer system unbootable, whether due to human error, hardware or software failure, natural disaster, and so on. In these cases it is most likely that the boot or system partition of the system is not available and the environment needs to be recovered before the standard restore operation can begin. This includes repartitioning and/or reformatting the boot partition and recovery of the operating system with all the configuration information that defines the environment. This has to be completed in order to recover other user data.

After a computer disaster has occurred, the system (referred to as the target system) is typically in a non-bootable state and the goal of Data Protector disaster recovery is to restore this system to the original system configuration. The difference between the affected and the target system is that the target system has all faulty hardware replaced.

Data Protector supports the following disaster recovery methods:

- Manual disaster recovery (UNIX only)

  This is a basic and very flexible disaster recovery method. You need to install and configure the DR OS. Then use Data Protector to restore data (including the operating system files), replacing the operating system files with the restored operating system files.

- Automated disaster recovery

  Automated System Recovery (ASR) is an automated system on Windows systems, which reconfigures a disk to its original state (or resizes the partitions if the new disk is larger than the original disk) in the case of a disaster.

- Disk delivery disaster recovery

  On Windows clients, the disk of the affected system (or the replacement disk for the physically damaged disk) is temporarily connected to a hosting system. After being restored, it can be connected to the faulty system and booted. On UNIX systems, the auxiliary disk, with a minimal operating system, networking, and Data Protector agent installed, is used to perform Disk Delivery Disaster Recovery.

- Enhanced Automated Disaster Recovery

  Enhanced Automated Disaster Recovery (EADR) is a fully-automated Data Protector recovery method for Windows clients and Cell Manager, where user intervention is reduced to minimum. The system is booted from the disaster recovery CD ISO image, and Data Protector automatically installs and configures DR OS, formats and partitions the disks, and finally recovers the original system with Data Protector as it was at the time of backup.

- One-Button Disaster Recovery

  One-Button Disaster Recovery (OBDR) is a fully-automated Data Protector recovery method for Windows clients and Cell Manager, where user intervention is reduced to a minimum. The system is booted from the OBDR tape and automatically recovered.

For lists of disaster recovery methods that are supported on different operating systems, see the latest support matrices at http://www.hp.com/support/manuals.

Single-pass disaster recovery

Data Protector delivers centralized system recovery (bare metal recovery) to virtual or physical servers (from P2V or V2P) from a single backup—at no additional cost. In addition, Data Protector streamlines the disaster recovery process. Unlike competitors, Data Protector enables customers to create a disaster recovery image from any existing full backup. IT staff simply check one box in the Data Protector console, and Data Protector makes sure that all of the necessary image information for a full system recovery is included in full backups. A DR image can then be created on a CD, DVD, USB key, or posted to an FTP site for download by remote or branch offices. Once the backup administrator initiates the disaster recovery process, Data Protector automatically rebuilds the system and the partitioning.

Distributed granular recovery

Data Protector Granular Recovery Extension empowers VMware and Microsoft SharePoint administrators to recover single items directly from the application administrator's console, without asking for assistance from the backup administrator. Data Protector shortens the single item recovery process by allowing application administrators to recover single items from Data Protector disk or tape backups.

Figure 13 shows the architecture of the Granular Recovery Extension for VMware. The Data Protector components are colored blue, and VMware components are black.

Figure 13: Data Protector Granular Recovery Extension for VMware

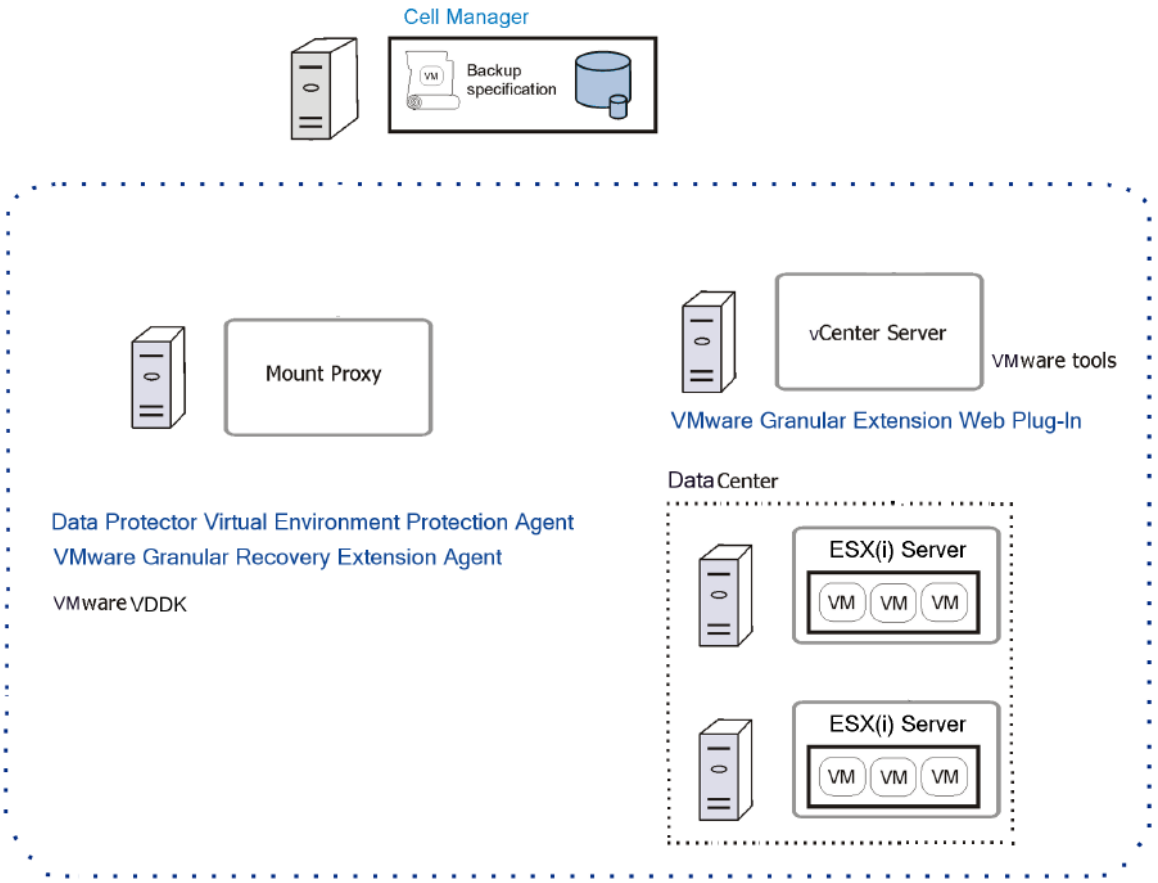
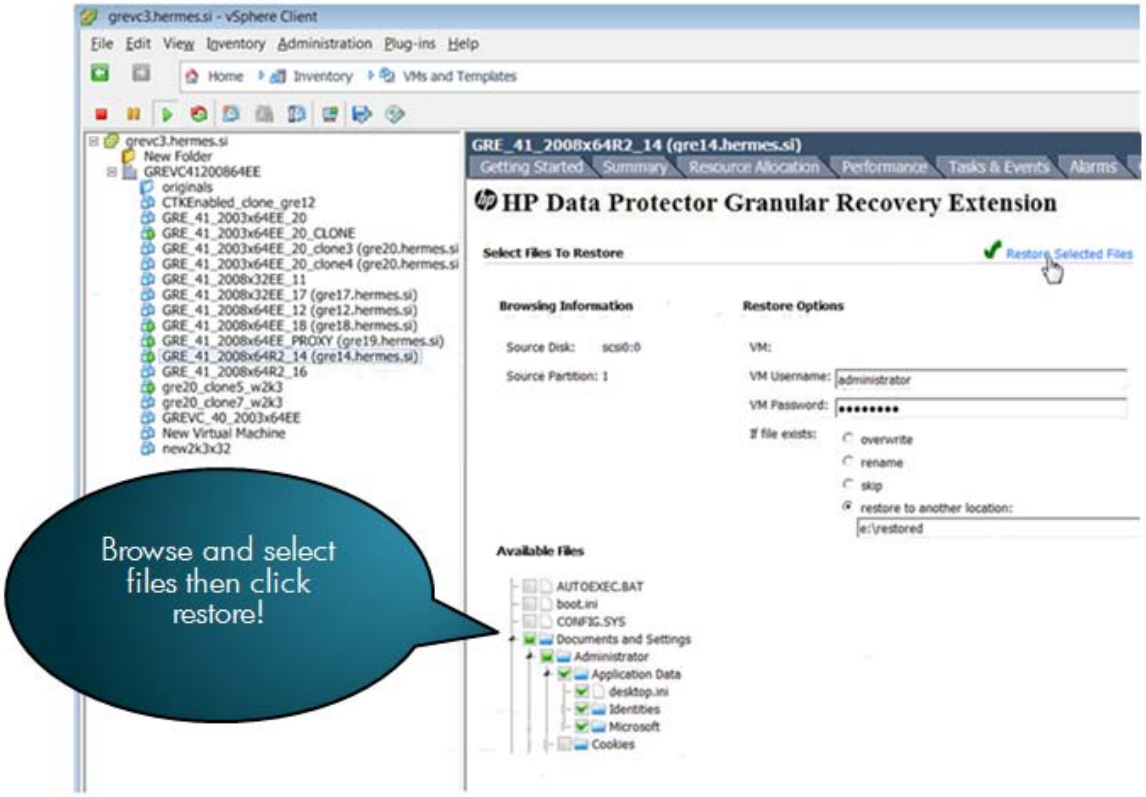


Figure 14 shows an example for VMware. VMware single items can be browsed, selected and restored by the VMware administrator.

Figure 14: Selecting VMware items with the Granular Recovery Extension



Synthetic and virtual full backups

Storage optimization needs can be met by implementing synthetic or virtual full backups, which can provide significant space reduction in space and are simple and cost-effective. For customers who are already using Data Protector Advanced Backup to Disk, synthetic and virtual full backup technology can be implemented without additional cost.

Synthetic full backup consolidates the most recent and all previous incremental backups into a new full backup to reduce impact on production file systems. Data Protector enables customers to easily copy a synthetic full backup from disk to tape, which can be stored on a single tape for disaster recovery. A full restore from a synthetic full backup is as fast as from a conventional full backup, since there is no need to retrieve data from incremental backups.

Virtual full backups offer the same benefits as a synthetic full backup without copying the data. Virtual full backup performs a full backup of file systems on disk using pointers to existing data for space savings and faster backup and recovery of file systems to and from disk. Customers can achieve up to 95% reduction in space when compared to a regular full backup.

Storage Area Networks

The Storage Area Network (SAN), shown in Figure 15, is a network dedicated to data storage and based on high-speed Fibre Channel technology. SAN provides off-loading storage operations from application servers to a separate network. Data Protector supports this technology by enabling multiple hosts to share storage devices connected over a SAN, which allows multiple-system to multiple-device connectivity.

When using Data Protector in the SAN environment, you have to consider the following:

- Each system can have its local or pseudo-local device, although the devices are typically shared among several systems. This applies to individual drives as well as the robotics in libraries.

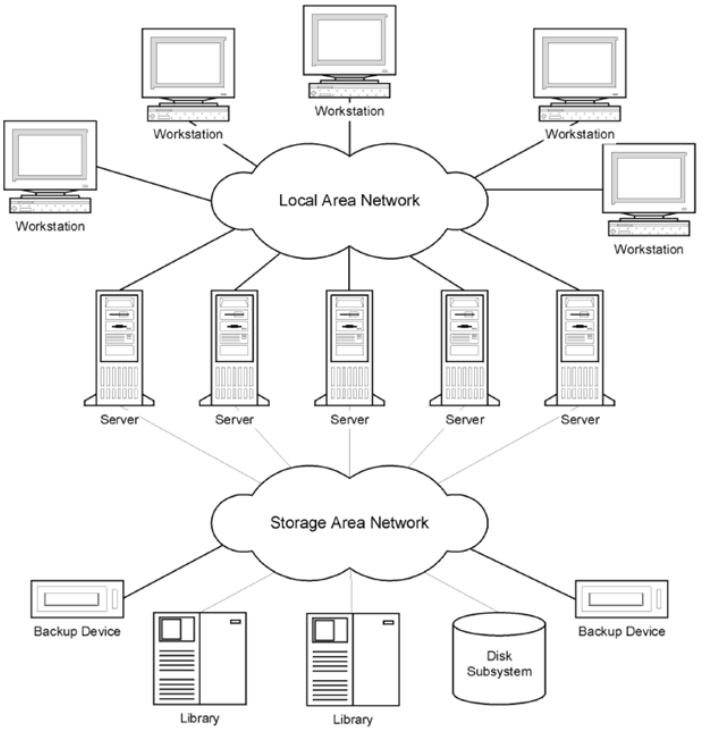
- You have to take care to prevent several systems from writing to the same device at the same time. The access to the devices needs to be synchronized between all systems. This is done using locking mechanisms.
- SAN technology provides an excellent way of managing library robotics from multiple systems. This creates the ability to manage the robotics directly, as long as the requests sent to the robotics are synchronized among all the systems involved.

Figure 15: Storage Area Network

– Network dedicated to data storage

– Based on high-speed Fibre Channel technology

– Enabling multiple hosts to share storage devices



Large file servers

Windows file servers could maintain very large file systems, in the worst case with millions of small files. In this case, backup and restore operations could take a very long time. Tests have shown that this is basically a Windows NTFS file system problem and not caused by the backup application. For instance, UNIX file systems provide a much better read and write performance.

Data Protector can improve the situation by reading data in parallel (concurrency). You can also consider using the Windows Change Log Provider for incremental backups, which will not scan the file system before backup.

The Windows NTFS Change Log Provider, based on the Windows Change Journal, queries the Change Journal for a list of changed files rather than performing a file tree walk. The Change Journal reliably detects and records all changes made to the files and directories on an NTFS volume, which enables Data Protector to use the Change Journal as a tracking mechanism to generate a list of files that have been modified since the last backup.

This is very beneficial for environments with large file systems, where only a small percentage of files change between backups. In this case, the process of determining changed files completes in a much shorter period of time.

Databases and applications

This section gives a brief description of the Data Protector integration with databases and applications. Data Protector supports Microsoft Exchange Server, Oracle Server, IBM DB2 UDB, and Informix Server, as well as with virtualization environments, such as VMware Virtual Infrastructure and Hyper-V.

For a detailed list of supported databases and applications, see the latest support matrices at http://www.hp.com/go/dataprotector.

Overview

From the user's perspective, a database is a set of data. Data in a database is stored in tables. Relational tables are defined by their columns and are given a name. Data is stored in rows in the table. Tables can be related to each other, and the database can be used to enforce these relationships. Data can thus be stored in relational format or as object-oriented structures such as abstract data types and methods. Objects can be related to other objects, and objects can contain other objects. A database is usually managed by the server (manager) process that maintains data integrity and consistency.

Whether you use relational structures or object-oriented structures, databases store data in files. Internally, these are database structures that provide a logical mapping of data to files, allowing different types of data to be stored separately. These logical divisions are called tablespaces in Oracle, dbspaces in Informix Server, and segments in Sybase.

Figure 16: Relational database concept

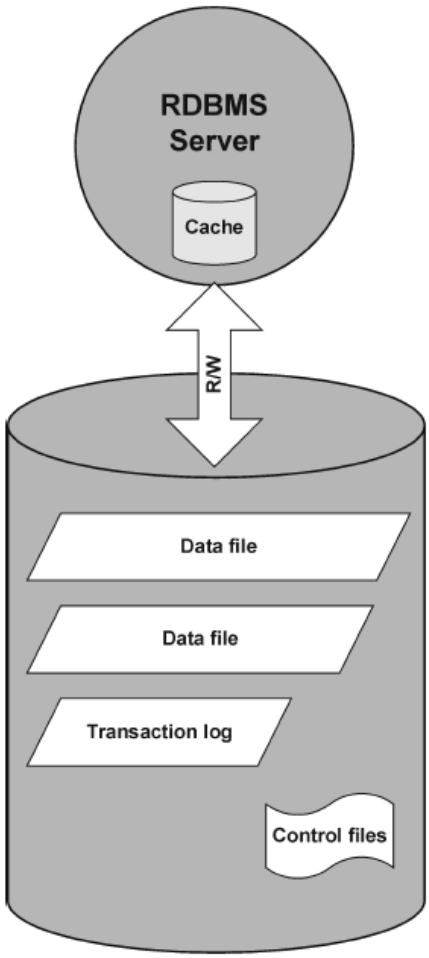


Figure 16 shows a typical relational database with the structures described below.

Data files are physical files that contain all of a database's data. They change randomly and can be very large. They are internally divided into pages.

Transaction logs record all database transactions before they are further processed. Should a failure prevent modified data from being permanently written to data files, the changes can be obtained from log files. Any kind of recovery is done in two parts: roll forward, which applies transaction changes into the main database and roll back, which removes uncommitted transactions.

Control files hold information about the physical structure of the database, such as, database names, names and locations of a database's data files and log files, and the time stamp of the database's creation. This control data is kept in control files. These files are critical for the operation of the database.

The cache of the database server process contains the most-often used pages of the data files.

The following is the standard flow of transaction processing:

1. A transaction is first recorded into the transaction log.
2. Changes required in the transaction are then applied to cached pages.
3. From time to time, sets of modified pages are flushed to data files on disk.

Filesystem backup of databases and applications

Databases are constantly changing while they are online. Database servers consist of multiple components that minimize response time for connected users and increase performance. Some data is kept in the internal cache memory and some in temporary log files, which are flushed at checkpoints.

Because data in a database can change during a backup, a filesystem backup of database files makes no sense without putting the database server into a special mode or even offline. Saved database files must be in a consistent state, otherwise the data is of no use.

The following steps are required to configure a filesystem backup of the database or application:

- Identify all data files.
- Prepare two programs, one to to shut down the database and and one to start it up.
- Configure the filesystem backup specification with all the data files included, and specify the shut-down program as a pre-exec command and the start-up program as a post-exec command.

This method is relatively simple to understand and configure, but has one key disadvantage: the database is not accessible during the backup, which is unacceptable for most business environments.

Online backup of databases and applications

To overcome the necessity of shutting down the database during a backup, database vendors have prepared interfaces that can be used to put databases temporarily into special modes to save the data to tapes. Server applications are thus online and available to users during the backup or restore process. These application-specific interfaces allow backup products, like Data Protector, to back up or restore logical units of the database application. The functionality of the backup APIs varies depending on the database vendor. Data Protector integrations are available for major databases and applications. For a detailed list of supported integrations, see the *HP Data Protector Product Announcements, Software Notes, and References*.

The essence of the backup interface is that it provides the backup application with consistent data (even if it may not be consistent on the disk) while at the same time keeping the database operational.

Figure 17: Data Protector integration with relational database

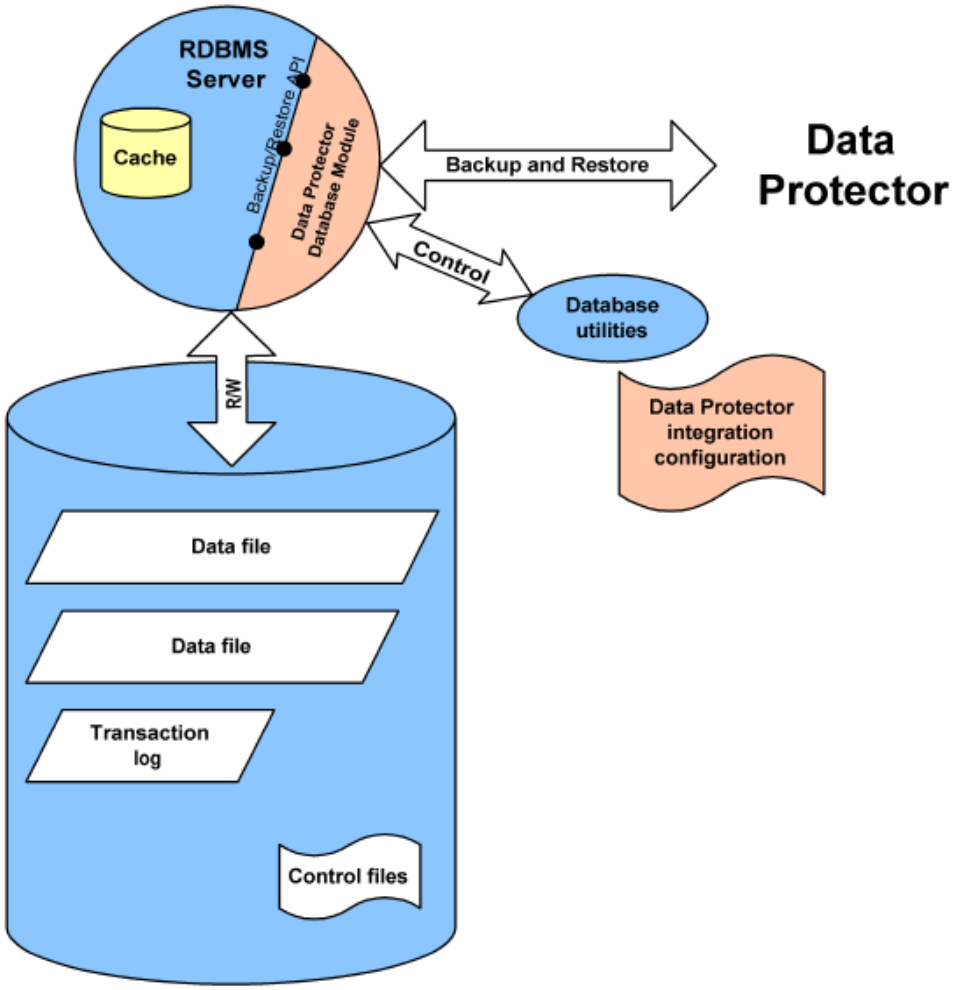


Figure 17 shows how a relational database is integrated with Data Protector. Data Protector provides a database module that is linked in to the database server. The database server sends data to Data Protector and requests data from it. Database utilities are used to trigger backup and restore operations.

A typical procedure to configure the backup of a database through the Data Protector integration is as follows:

1. A database/application-specific agent is installed on the database system
2. The Data Protector integration is configured for each database. Data needed for Data Protector to work with this database is stored on the database system (into configuration files or registry entries). Typically, this includes pathnames, and user names and passwords.
3. The backup specification is prepared using the Data Protector user interface.

Besides the key advantage of the database being online all the time, there are also other benefits of using the Data Protector integrations with the databases:

- There is no need to specify the location of data files. These can be located on different disks.
- The logical structure of the database can be browsed. It is possible to select only a subset of the database.
- Applications are aware of the backup operation, and keep track of which parts are backed up.
- Several modes of backup are possible. Besides full backups, users can select (block-level) incremental backups, or only the backup of transaction logs.
- Several modes of restore are possible, and after the restore of data files, the database can automatically restore transaction logs and apply them as configured.

Virtualized environments

This section gives a brief description of the Data Protector integration with virtualization environments, such as VMware Virtual Infrastructure and Hyper-V.

Offline backup

If you perform an offline backup of a VM, you must shut down or suspend that VM. If it is shut down or suspended, users cannot access any of its applications. This is not useful for today's critical 24/7 operations. However, if the enterprise has its own homegrown application that does not have a specific backup interface that can be used, you might be forced to shut down the VM to perform a backup.

Tradional backup

With the traditional backup method, an online Backup Agent is installed inside the VM as shown in Figure 18. This is probably the most common method, and it is the way most enterprises protect servers and applications in the physical world. With Exchange, SharePoint, Oracle, or SAP, administrators would put an online Backup Agent in the physical server that can communicate with the application to perform a scheduled backup or restore. In the virtual world, it is easiest for most to use this method; simply put an online Backup Agent inside the VMs.

Figure 18: Backup inside a VM

However, there is a tradeoff. Backup negatively impacts all VMs and the physical hosts. This is easily envisaged by thinking about undertaking backup in the physical world. Administrators typically perform backup from 20:00 hrs in the evening until 6:00 hrs in the morning. This period of time is usually referred to as the backup window, and backup is undertaken at a time when very few users are online. After it begins, it consumes vast quantities of resources on the physical host. It draws on memory, hard disk, I/O, and the CPU. Users would not be able to access applications if this backup occurred during the working day.

In the virtual paradigm, if the administrators start backing up VMs at the same time, it puts a massive load on the host and degrades performance. A VMware DRS may try to redistribute the load. However, this is a problem DRS simply cannot solve. That is why backup has been one of the primary challenges in a virtual environment; how do administrators properly back up virtual environments? VMware has addressed this challenge with the introduction of the vStorage API for Data Protection.

VMware vSphere and vStorage API for Data Protection

The VMware vSphere and vStorage API for Data Protection (VADP) is not a backup and recovery product. It does not back up data to disk or tape, nor recover data from disk or tape. However, it serves as an important platform from which backup software applications, such as Data Protector, can create recoverable, application-consistent backups. The VADP does away with the crash consistent state for Microsoft VMs (for instance machines running the Microsoft OS and applications that support Volume Shadow Copy Service (VSS) from Microsoft). VSS is Microsoft's tool for backing up its applications; VADP integrates with that tool.

VADP enables Data Protector to perform agent-less, image-level, full, incremental, and differential backups.

Data Protector Virtual Environment Agent

Integrating Data Protector into the VMware infrastructure is straightforward, and requires configuration with just a few components, including the Data Protector Cell Manager and the Virtualization Environment Agent (VEAgent).

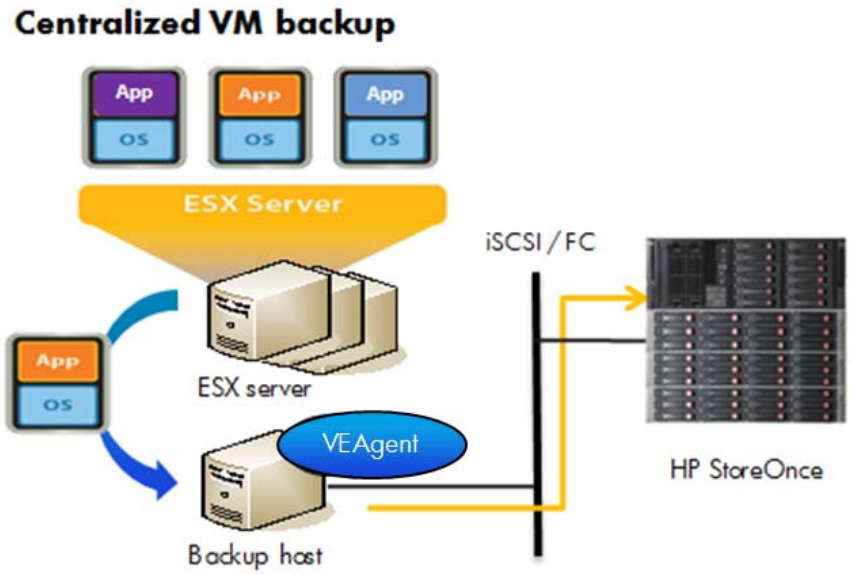
The Data Protector Cell Manager is the system that manages the media and details of what was backed up as well as what is available for restore, and much more. The Data Protector VEAgent, shown in Figure 19, is a fully integrated virtualization agent for both VMware and Microsoft Hyper-V environments. The VEAgent can easily be

installed via the integrated installation solution from Data Protector. The system on which this VEAgent is installed becomes the "backup host" and manages the data transfer to the target devices for backup and restore.

Finally, IT staff can import Data Protector into either a vCenter Server or an ESX server. This simplifies management tasks by leveraging the vCenter central administrative database, removing the need to manage ESX servers one by one.

As mentioned above, the VEAgent support both VMware and Hyper-V. This combined virtualization agent helps simplify backup operations in the enterprise environment.

Figure 19: Data Protector Virtual Environment Agent (VEAgent)



Simplifying snapshots

The Data Protector Zero Downtime Backup extension effectively reduces the physical or virtual backup window in any physical or virtual environment, allowing administrators to take advantage of array-based snapshots to back up the physical or virtual environment at any time of the day or night with zero impact to the production environment. This allows administrators to achieve any RPO and RTO for any virtual application. In VMware environments, all applications running as VMs can be protected as frequently as required, because all backups are processed by the storage array. Data Protector creates a snapshot (sometimes referred to as a replica or a clone) and moves that copy to a disk or tape for long-term storage. In addition, it can even be left on the disk for rapid recovery.

Data Protector helps administrators employ a "set it and forget it" backup policy, and helps ensure protection in even the most dynamic environments, both physical and virtual. One click in the console and Data Protector will automatically apply snapshot configuration policies to new VMware VMs before the backup session begins.

Data Protector provides snapshot support for both HP and non-HP arrays, including the HP StorageWorks P2000, P4000, EVA, and P3000, HP 3PAR, EMC CLARiiON, EMC Symmetrix DMX, and Network Appliance.

Down-to-the-second recovery

The Data Protector Instant Recovery extension can recover snapshots to any point in time—down to the exact second specified by the backup administrators—all from a single console. Data Protector takes advantage of what is called "roll forward functionality."

For example, a snapshot-based backup is scheduled to occur at 06:00 hrs using zero downtime backup. At 07:30 hrs, an error occurs (a database is accidentally deleted, a virus is introduced in the messaging system, or corruption occurs at some level). Administrators could recover the snapshot taken at 06:00 hrs, but would lose roughly 90 minutes of data. In some environments that might be acceptable. But in most mission-critical environments, losing 90 minutes worth of data is no longer tolerated.

Data Protector Zero Downtime Backup and Instant Recovery extensions provide protection without compromise—any point-in-time recovery with a roll-forward capability. It is a faster recovery method with a higher RPO and

RTO. It meets the most demanding SLAs, and protects not only vSphere and VI3 but also Microsoft Hyper-V, physical servers and other virtual environments.

Recommendations for VMware data protection

Depending upon the environment, there are various recommendations for VMware data protection as shown in table 1:

Table 1: Recommendations for virtualized environments.

Method	Best for	Advantages	Disadvantages	Cost	Comment
Traditional method	Small VMware environments	• Same as physical backup method	• Performance • No enablement of advanced VMware features, such as Change Block Tracking	$	Data Protector leveraging Disk Agent inside virtual machine, Disk Agent-based filesystem backup does not require a license.
Data Protector leveraging vStorage API	Larger virtual environments	• Single management for physical and virtual servers • Full range of backup media and targets	• Cannot select single directories or files	$$	Single file restore options available
Data Protector ZDB/IR with HP disk arrays	Customers with business mission-critical applications —who require instant recovery	• Application integration • No backup load on virtual machines • Recovery time of minutes	• Only supported on specific storage hardware	$$$	Fast recovery times with full data protection and archive options

For small VMware environments, a traditional data protection process, with a backup and recovery software agent installed in each virtual machine to manage network backup to a shared backup device may be adequate.

For larger VMware vSphere installations running mid-range applications with a mix of applications running on physical servers and virtual machines, leverage the vStorage API to provide various levels of integration with VMware as well as deep application integration.

To meet the needs of organizations without backup windows and applications that must run continually, HP offers Data Protector zero downtime backup (ZDB) and instant recovery (IR) with HP disk array snapshot capabilities.

Virtualization technology such as VMware enables organizations of all sizes to slash the number of physical servers deployed with the intention of trimming costs, cutting complexity, increasing operational flexibility, and reducing power and space requirements. The data protection requirements of VMware installations are similar to physical environments in terms of being driven by required recovery time (RTO) and recovery point objectives (RPO). Although data protection processes developed for physical environments can be used, processes designed for VMware environments are more flexible and scalable and do not negate the benefits derived from using a virtualized infrastructure.

Data Protector supports several virtualization applications such as VMware, Microsoft Hyper-V and Virtual Server. For a detailed list, see the latest support matrices at www.hp.com/support/manuals.

For more details, see the *HP Data Protector Integration Guide for Virtualization Environments*.

Disk-based backup and virtual tape libraries

This section describes the HP StorageWorks VLS and D2D systems including the integration with Data Protector.

In disk solutions, data is backed up from an application server (disk) over a dedicated SAN to a disk-based system and from there to a traditional tape library. This provides enhanced solutions for slow servers, single-file restores, and perishable data.

One of the particular benefits of the VLS and D2D is that they make a disk array appear like tape library to your backup server. Implementation requires no new software and no significant redesign of your backup processes.

You can optimize your backup environment with VLS and D2D if you are:

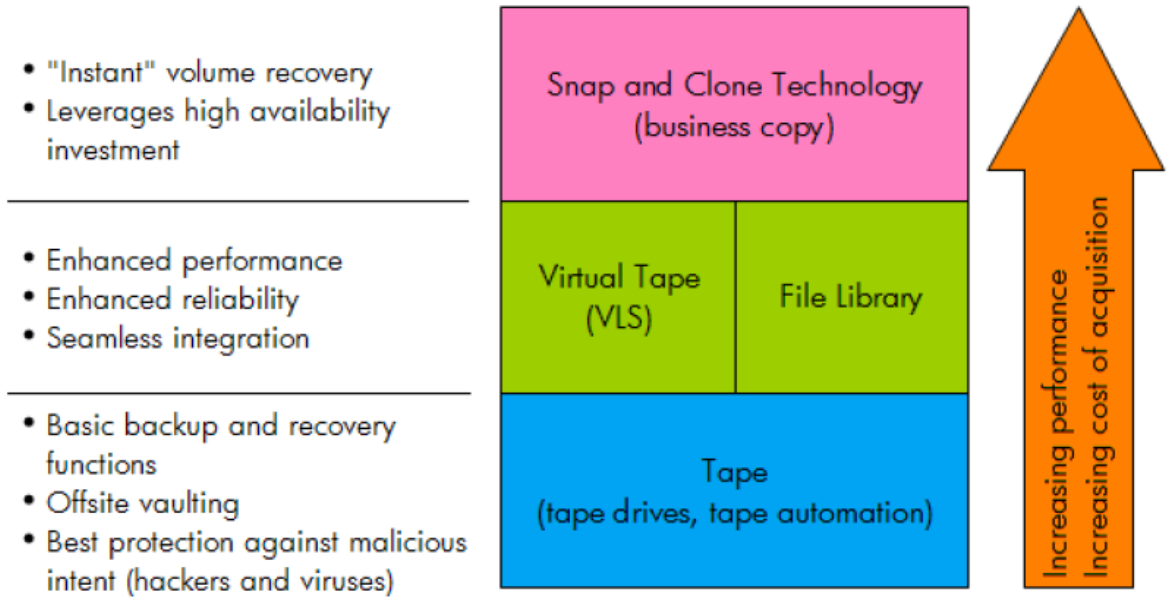
- Not meeting backup windows due to slow servers.
- Not consistently streaming your tape drives.
- Dealing with restore problems caused by interleaving.
- Performing many restores (such as single file or small database restores).
- Backing up data that has a short life.
- Having issues with backup reliability.
- Using snapshot and clone technology for non-critical data (which makes the storage inappropriately expensive for the nature of the data).
- Looking to de-emphasize tape in your environment. Bear in mind that removable media remains valuable in its own right and for particular purposes such as site protection and protection from malicious attack (for example, viruses and hackers), data distribution, data copy, archive, and regulatory compliance.
- Improving media management. You can keep incremental backups on virtual tape and send full backups straight to tape.

Note: Tape holds its value for ease of vaulting, economical long-term retention, and immutability (with WORM). It is the last step in your data's storage cycle.

Where virtual tape fits in the big picture

Virtual libraries are not necessarily the only feature of your backup plans, but they can be an integral part of a successful solution. Figure 20 illustrates the common backup technologies and their relative benefits and costs.

Figure 20: Benefits of common backup technologies



Impact of the Virtual Library System

The VLS is the solution for the following problems:

1. Not meeting the backup window

   Some administrators run the risk of losing data, often caused by increased data volumes and inadequate backup hardware. This also impacts the production environment though accepting lower performance until backup jobs have completed. In many companies, nightly backup reports are delivered up the IT management infrastructure. Excessive backup failures (either on success rate or SLA basis) have visibility.

   With VLS, backup performance is accelerated by allowing more jobs to happen at the same time so the backup window is achieved. It improves process reliability (by removing tape handling, bad media, and physical drive related errors), and removes infringement of the backup jobs on the production resource. It also reduces failures and SLA violations resulting in fewer reportable incidences.

2. Slow restores

   Storage administrator resources are devoted to this task at the expense of other activities (such as development) by accepting lower productivity in the department that is waiting on the resources.

   With the VLS, restore times are reduced, which in turn reduces the load on storage administrators and help desk operators, resulting in more time to do application development. It allows resources that were affected by data loss to reach productivity quicker, and results in a more productive workforce.

3. Inefficient media usage

   IT departments purchase more media and are over-paying for offsite services, since more media than is necessary goes offsite and the subscription and retrieval costs are higher than necessary.

   With the VLS, the number of tape copies that are required is reduced. Many users are cloning tapes—one for on-site, one for off-site. This eliminates the on-site media expense (on-site media expense is recurring due to media wear-out). Media are more efficiently filled, resulting in fewer pieces needed. Some users are not fully filling media for various reasons (complexities with multiplexing streams from the LAN, short backups such as database redo logs). Furthermore, copy jobs can de-multiplex the virtual media to physical media and put data back-to-back, resulting in better usage.

4. Perishable solutions usage

   IT departments rip and replace solutions when responding to unplanned growth and changes. VLS scales very well, allows flexible device configurations, and offers advanced features for data deduplication (replication and automigration).

The impact summary of the VLS is shown is shown in table 2:

Table 2: Impact summary of the Virtual Library System

Problem	What people do today	How the HP Virtual Library System solves the problem
Not meeting backup window	• Gamble • Accept lower performance during backup jobs • Report backup failure to management	• Accelerates backup performance • Improves backup process reliability
Slow restores	• Devote storage administrator resources • Accept lower productivity	• Reduces workload of storage administrators and helps desk operators • Allows quicker time to productivity by affected resources
Inefficient media usage	• Purchase more media • Over-pay for offsite services	• Reduces number of tape copies • Fills media more efficiently
Perishable solutions (becoming too small, failing to work with new applications, and so on)	• Rip and replace solutions • Deploy point solutions for the worst problems	• Scales capacity and throughput independently • Uses data deduplication, deduplication-enabled replication and automigration

Typical VLS environments

In a typical enterprise backup environment, there are multiple application servers backing up data to a shared tape library on the SAN. Each application server contains a remote backup agent that sends the data from the application server over the SAN fabric to a tape drive in the tape library. However, because backup over the SAN is single-threaded (a single host is backing up to a single tape drive), the speed of any single backup is likely to be limited. This is particularly true when the environment has high-speed tape drives such as Ultrium 5. The hosts simply cannot keep the drives streaming at capacity.

Note: HP Ultrium drives will adjust the tape speed to match the data stream to prevent "back-hitching." However, the tape drive is still not operating at optimal performance and cannot share bandwidth with another backup job.

Enterprise data centers with slow SAN hosts in the environment may be unable to use the full performance of high-speed tape drives. Also, shared tape libraries on the SAN can be difficult to configure both in the hardware and in the data protection software.

Typical D2D environments

In a typical entry-level or mid-range backup environment, the backup application is performing LAN backups to a dedicated (non-shared) backup target, such as a tape library, connected to the single backup server. Multiple instances of the backup application will generally each require their own dedicated backup target. These environments may also be remote branch offices, each with their own local backup application.

As with the VLS, the backup speed of a single host backing up to a single tape drive is normally limited by the host (which cannot stream high-speed tape drives such as Ultrium 5), so currently tape backups use multiplexing to interleave multiple hosts' backups together to a single tape drive, which impacts restore performance. The addition of a D2D device to these environments allows de-multiplexing of the backups so that restore performance is improved. The deduplication allows for a longer retention time on disk without needing significantly higher disk capacities, and the deduplication-enabled replication allows cost-effective off-site copying of the backups for disaster protection.

Alternatives

Alternatives to virtual tape solutions include:

- Physical tape
- NAS (network attached storage)
- Application-based disk backup (disk to disk, backup to disk, disk to disk to tape)
- Business copy (snapshot and clone solutions)

Physical tape is the foundation for data protection and should be a part of most data protection solutions (except those with highly perishable data).

Consider a direct-to-tape scheme if:

- You are doing large image backups (such as databases).
- Your servers can stream the tape drives.
- You do not need fast single-file restore.

An NAS device acting as a backup target (via NFS or CIFS network file system protocols) is an alternative to a virtual library. However, this protocol has significant performance and scaling limitations; writing backups over TCP/IP and NFS/CIFS to the NAS target uses much more CPU on the backup infrastructure compared with Fibre Channel SAN. In addition, an NAS mount point does not scale to the size of an enterprise virtual tape library. For example, a VLS can present a single virtual library target containing multiple petabytes of tape capacity with all backup jobs configured to use the one common shared high-performance high-capacity VLS backup device.

Consider an NAS target if:

- You do not have high performance requirements.
- You do not want to run SAN backups.
- You do not need the backup target to significantly scale capacity or performance.
- You want to run Data Protector Virtual Full Backups.

Application-based disk backup utilizes the file library functionality of Data Protector, which is good for small or isolated jobs. When a large-scale implementation is required, virtual tape offers a more easily managed and higher performing solution.

Consider a file library device if:

- The application is in a LAN or LAN/SAN hybrid configuration.
- Fewer than four servers write data to secondary disk storage.
- You can redeploy existing arrays as secondary disk storage.
- Your environment is static.

The concept of application-based disk backup is shown in Figure 21. It uses a low-cost array as backup target (black arrow from host to array) and also as source for a disk-to-tape copies (colored arrows from array to tape). This concept is also known as D2D2T (Disk to Disk to Tape).

Figure 21: Application-based disk backup (file library)

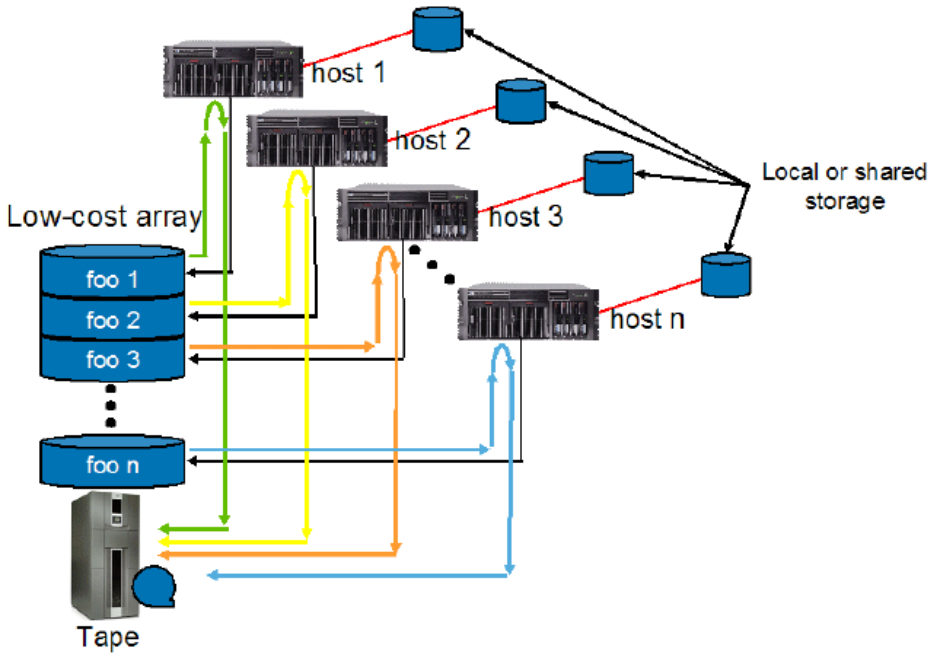


Table 3 compares application-based write-to-disk with virtual tape devices. Virtual tape devices require less setup and management effort, offer data compression, and good performance. However, the acquisition of virtual tape devices costs more than application-based write-to-disk.

Table 3: Comparison of application-based write-to-disk with virtual tape devices

Factor	Virtual tape devices	Write-to-disk
Setup and management complexity	• Sets up just like a physical tape library	• Requires configuration of RAID groups, LUNs, volumes, and filesystems
Data compression	• Software- or hardware-enabled (software compression generally decreases performance)	• No device-side data compression available
Performance	• Hardware devices are tuned for sequential read and write operations	• Performance dependent on target array or server
Cost	• Greater acquisition cost • Backup software licenses as if a physical library, or per TB • Storage efficiency gained through compression • Lower management overhead	• Free or licensed per TB in most backup applications • Higher management overhead

Business-copy solutions (array snapshots and clones) generally involve a much higher cost than a virtual library system. You might, however, implement such a solution if:

- Virtually instant recovery is critical.
- You need to leverage a high-availability investment.
- You are performing image recovery rather than file recovery.
- You need a zero downtime solution

Deduplication

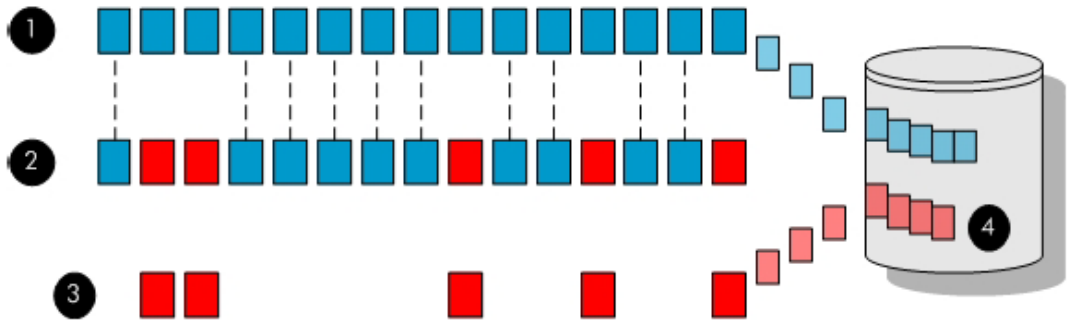
In recent years, the amount of data that companies produce has been steadily increasing. To comply with government regulations, or simply for disaster recovery and archival purposes, companies must retain more and more data. Consequently, the costs associated with data storage—labor, power, cooling, floor space, and

transportation of physical media—have all risen. Virtual tape libraries have become a cornerstone in modern data protection strategy due to their many benefits. Chief among these is cost. The list of virtual tape benefits also includes seamless integration into existing backup solutions, improved SAN backup performance, and faster single file restores than those performed with physical tape.

Deduplication, one of the most significant storage enhancements in recent years, promises to reshape future data protection and disaster recovery solutions. This technology is ideal for virtual tape libraries.

Deduplication technology, shown in Figure 22, references blocks of data that have been previously stored, and only stores new backup data that is unique. Data that is not unique is replaced with a pointer to the location of the original data. Because there is often a great deal of duplicate data present from one backup session to the next, disk space is consumed by similar or identical iterations of data. Deduplication greatly improves storage efficiency by only storing an instance of data once, while still allowing backup streams to be restored as if they had been retained in their entirety.

Figure 22: Unique backup data



Item	Task
1	Data from the first backup stream is stored to disk
2	Duplicate data (in blue) as well as unique data (in red) in a second backup stream is identified
3	Duplicate data in the second backup stream is eliminated
4	Unique data in the second backup stream is stored to disk

HP StorageWorks deduplication solutions

HP offers two deduplication technologies:

- **HP Accelerated deduplication**, a licensed feature available with HP StorageWorks Virtual Library Systems (VLS).
- **HP StoreOnce deduplication**, an integrated feature with HP StorageWorks D2D Backup System.

Both HP deduplication solutions offer the following benefits:

- Longer retention of data.
- Faster, less expensive recoveries and improved service levels.
- Fewer resources consumed, reducing operational costs.
- Completely transparent to host.
- No data is lost—backup streams can be fully restored.
- Block or chunk level deduplication, providing greater reduction of data.
- Even greater reduction of data when combined with traditional data compression.

HP Accelerated deduplication and HP StoreOnce deduplication are designed to meet different needs, as shown in table 4:

Table 4: HP deduplication solutions

HP Accelerated deduplication	HP StoreOnce deduplication
• Intended for enterprise users • Uses object-level differencing technology • Fastest possible backup performance • Fastest restores • Most scalable solution in terms of performance and capacity • Potentially higher deduplication ratios	• Intended for mid-sized enterprise and remote office users • Uses hash-based chunking technology • Integrated deduplication • Lower cost and a smaller RAM footprint • Backup application and data type independence for maximum flexibility

The storage capacity saved by deduplication is typically expressed as a ratio, where the sum of all pre-deduplicated backup data is compared with the actual amount of storage the deduplicated data requires. For example, a 10:1 ratio means that ten times more data is being stored than the actual physical space it would require.

The most significant factors affecting the deduplication ratio are:

- How long the data is retained
- How much the data is changed between backups

Table 5 provides an example of storage savings achieved with deduplication. However, many factors influence how much storage is saved in your specific environment. Based on the retention policies shown below, six months of data without deduplication requires 12.75 TB of disk space. With deduplication, six months of data requires less than 1.25 TB of storage.

Retention policy:

- 1 week, 5 daily incremental backups
- 6 months, 25 weekly full backups

Data parameters:

- Data compression rate = 2:1
- Daily change rate = 1% (10% of data in 10% of files)

Table 5: 1 TB File server backup

	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**
Approximately 11:1 reduction in data stored		

Table 6 is an example that may not reflect the savings that all environments achieve using deduplication. As shown, deduplication ratios depend on the backup policy and on the percentage change between backups.

Table 6: Deduplication ratio impact

Daily change rate	Backup policy					
	Daily full and weekly full			Daily incremental (10%) and weekly full		
	4 months *	6 months	1 year	4 months *	6 months	1 year
0.5%	15:1	19:1	25:1	12:1	16:1	23:1
1.0%	12:1	13:1	16:1	10:1	11:1	15:1
2.0%	8:1	9:1	9:1	7:1	7:1	9:1
*4 months = 5 daily + 17 weekly backups						

VLS and D2D deduplication is target-based; the process runs transparently inside the hardware. This means that when the data is read (by copying to physical tape, restoring a backup, and so on), the device rebuilds the data. The data that is read is identical to the data that was originally written (like tape drive compression); there are no pointers in the read data.

Replication

Deduplication can automate the off-site process and enable disaster recovery by providing site to site deduplication-enabled replication at a lower cost. Because deduplication knows what data has changed at a block or byte level, replication becomes more intelligent, and transfers only the changed data instead of the complete data set. This saves time and replication bandwidth, and is one of the most attractive features that deduplication offers. Replication enables better disaster tolerance with higher reliability but without the operational costs associated with transporting data off-site on physical tape.

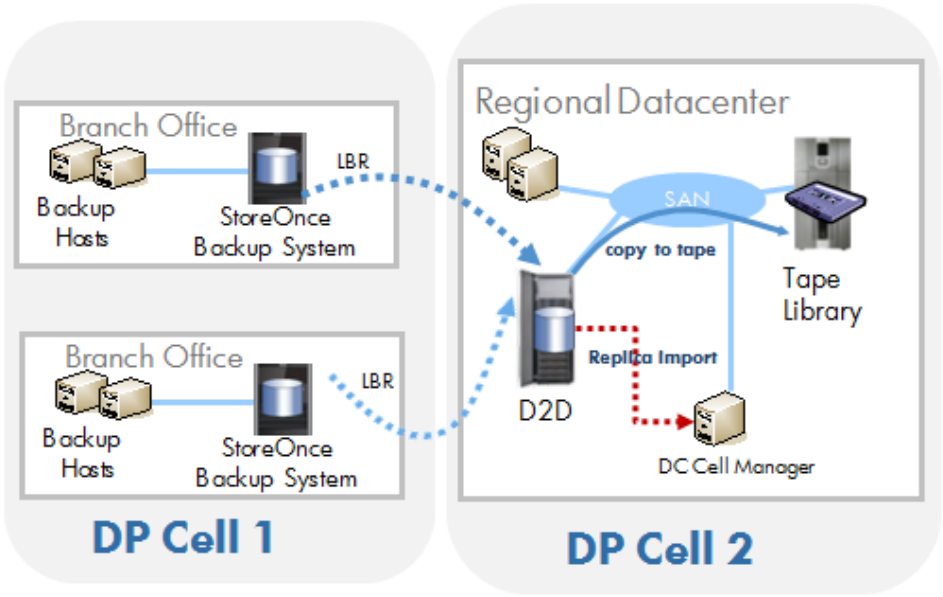
Replication provides end-to-end management of backup data from the small remote office to the regional site, and finally into the primary data center, all controlled from the primary data center, while also providing local access to backup data. Note that replication is within device families (VLS to VLS, D2D to D2D).

Data Protector is fully replication-aware, and, together with HP StoreOnce D2D and Virtual Library Systems (VLS) devices, provides a single management console from which to oversee deduplication-enabled replication between sites—for locally or geographically distributed environments.

Geographically-distributed organizations can take control of the data at its furthest outposts and bring it to the data center in a cost-effective way.

Figure 23 shows an example with the HP StorageWorks D2D which provides deduplication-enabled, low-bandwidth replication (LBR) for both VTL and NAS devices.

Figure 23: Branch office protection with low bandwidth replication



Replication provides a point-in-time mirror of the data on the source D2D (such as a branch office) at a target D2D system on another site (such as a regional data center). This enables quick recovery from a disaster that has resulted in the loss of both the original and backup versions of the data on the source site. Replication does not however provide any ability to roll-back to previously backed-up versions of data that have been lost from the source D2D. For example, if a file is accidentally deleted from a server and therefore not included in the next backup, and all previous versions of backup on the source D2D have also been deleted, those files will also be deleted from a replication target device, because the target is a mirror of exactly what is on the source device. **This is one good example why data should be copied to tape.**

Note: One of the most important aspects in ensuring that a replication will work in a specific environment is the available bandwidth between replication source and target D2D systems. In most cases a WAN link will be used to transfer the data between sites unless the replication environment is all on the same campus LAN. It is advisable that the HP StorageWorks Sizing Tool (http://www.hp.com/go/storageworks/sizer) is used to identify the product and WAN link requirements, because the required bandwidth is complex and depends on the following:

- Amount of data in each backup
- Data change per backup (deduplication ratio)
- Number of D2D systems replicating
- Number of concurrent replication jobs from each source
- Number of concurrent replication jobs to each target

What target device (emulation) should be used with a StoreOnce Backup System?

Data Protector supports Virtual Tape (VTL) and NAS, which have the following considerations:

VTL
- The existing backup environment uses tape.
- There is minimal impact on the backup methodology to change to the StoreOnce Backup System as primary target.
- It is easy to import media at the replication target from the Data Protector GUI (without scripts).
- Data Protector capacity-licensing enables flexible deployment without drive/slot licensing constraints.

NAS
- The customer is more familiar with backing up to a disk target, and for example, has never used tapes.
- The Data Protector File Library is used to back up to an NAS share.
- There is a need to implement sophisticated backup methods, such as Enhanced Incremental and

Synthetic Full backups.

- It provides network-based backup from systems that do not support iSCSI.

Note: VTL and NAS emulations can be created on the same appliance.

What backup type should be used with a StoreOnce Backup System?

- **Full**
  - Fastest recovery
  - Does the backup window allow this?

  Full gives the fastest recovery since there is only one virtual cartridge image set to restore, but can your backup window allow this backup to complete every day without impacting operations? This method transfers the most data over the course of the backup cycle.

- **Full + Incremental**
  - Faster backups between full backups
  - What is the data change rate?
  - Retention time of incremental same/different to full?

  Full + Incremental is what many customers use with physical tape media, and they are familiar with the strategy. However, a restore typically use many tapes or backup sessions, depending on where the restore point is with respect to the last full backup, for example, how many intervening incremental backups there are. This method transfers the least data (except for the special case Incremental Forever) during the backup cycle.

- **Full + Differential**
  - Compromise between 'Full' and 'Full + Incremental'
  - Less important for backup to disk

  Full + Differential has been used with physical tapes, because at most only two media sets are required for a restore. This backup method is a compromise, with increasingly more backup data during the backup cycle but fewer media required for restore.

- **Incremental Forever**
  - First backup is the only full backup
  - All subsequent backups are incremental

  Incremental Forever backup starts with one complete full backup and continues with incremental backups only. This transfers the least amount of data, at the expense of requiring many more media for a complete restore, especially as you get further in time from the original full backup. A further issue is that if only *one* incremental backup media is missing or damaged, the restore chain becomes broken at that point in time, and you really need to start with another full backup to adequately protect your data.

- **Enhanced Incremental**
  - Data Protector feature to add efficiency to incremental backups
  - Required for Synthetic Full and Virtual Full Backups
  - Requires a NAS target (not VTL)

  Enhanced Incremental is a Data Protector option that uses either internal file tracking techniques or the Windows Change Log Service to identify precisely which files have changed without having to walk the directory tree for each incremental backup. Tree walks are time-consuming for huge file systems with millions of files. Enhanced Incremental backups require a disk backup target, so they can only be used with a native file system disk or a NAS target.

- **Consolidating Data Protector Enhanced Incremental backups**
  - Synthetic Full (supported with a StoreOnce Backup System)
  - Virtual Full (**not advisable** with a StoreOnce Backup System)

  Enhanced Incremental backups can be consolidated to synthetically produce a full backup image from an earlier full and intervening incremental backups. Synthetic Full backups are supported with the StoreOnce Backup System, since each backup session produces a "media" set on disk. *Virtual Full* backups are **not supported** with the StoreOnce Backup System, since they rely on a more complex Distributed File Media Format (DFMF) disk layout, which produces a large number of files on the StoreOnce Backup System which in turn affects performance and the file limit per NAS share.

Network Data Management Protocol

Network Data Management Protocol (NDMP) is an open standard protocol for enterprise-wide backup of Network Attached Storage (NAS).

NAS concept

NAS is file-level computer data storage connected to a computer network providing data access to heterogeneous network clients.

NAS has the following characteristics:

- It is not designed to be a general purpose server, although it may technically be possible to run other software on a NAS unit. For example, NAS units usually do not have a keyboard or display, and are controlled and configured over the network, often using a browser.
- A fully-featured operating system is not needed on a NAS device, so often a stripped-down or customized standard operating system is used, such as Linux.
- NAS systems contain one or more hard disks, often arranged into logical, redundant storage containers or RAID arrays (redundant arrays of inexpensive/independent disks).
- NAS removes the responsibility of file serving from other servers on the network.
- NAS uses file-based protocols such as NFS (popular on UNIX systems), SMB/CIFS (Server Message Block/Common Internet File System—used with MS Windows systems), or AFP (used with Apple Macintosh computers).
- NAS units rarely limit clients to a single protocol.

NAS backup

The backup of data residing on a filer can now be done in two different ways:

1. On the Application Server
2. Via NDMP direct from filer to tape

The first case is a normal backup, using the Data Protector Disk Agent, which is installed on each application server. During a backup, data is transferred via the LAN to the system to which a tape drive is connected.

The second way uses the Data Protector NDMP integration, and performs the backup locally on the NDMP server.

This has two major advantages:

- The backup data is not transferred via LAN.
- There is no performance degradation on the application server.

NDMP concept

The NDMP (Network Data Management Protocol) concept is summarized as follows.

- Open standard protocol for enterprise-wide backup of heterogeneous network-attached storage (NAS)
- Specifies the communication between storage appliances and the backup software
- Developed by Network Applicance (NettApp) and IntelliGuard Software (acquired by Legato)
- Submitted to IETF (Internet Engineering Task Force)
- SNIA (Storage Networking Industry Association) has formed a group working towards IETF standardization of NDMP

The NDMP protocol creates a *universal agent* that can be used by NDMP-compliant and centralized backup applications. Communication takes place between the centralized backup application and this universal agent, which is provided by the network-attached storage.

The NDMP protocol was developed and implemented by Network Appliance and IntelliGuard.

The current list of NDMP-supported products includes backup software such as Data Protector, Messaging Appliance (Mirapoint) and NAS hardware (EMC, Network Appliance, Hitachi Storage Systems, and so on).

The full specification for NDMP was submitted to the Internet Engineering Task Force (IETF) in October 1996. The goal is that NDMP v4 becomes an IETF-accepted RFC (Request For Comment). To accomplish this, a group came together in SNIA (Storage Networking Industry Association) to work on this proposal to assure that the result,

NDMP v4, becomes a mechanism that is both necessary and sufficient for building data management solutions that address current and future needs.

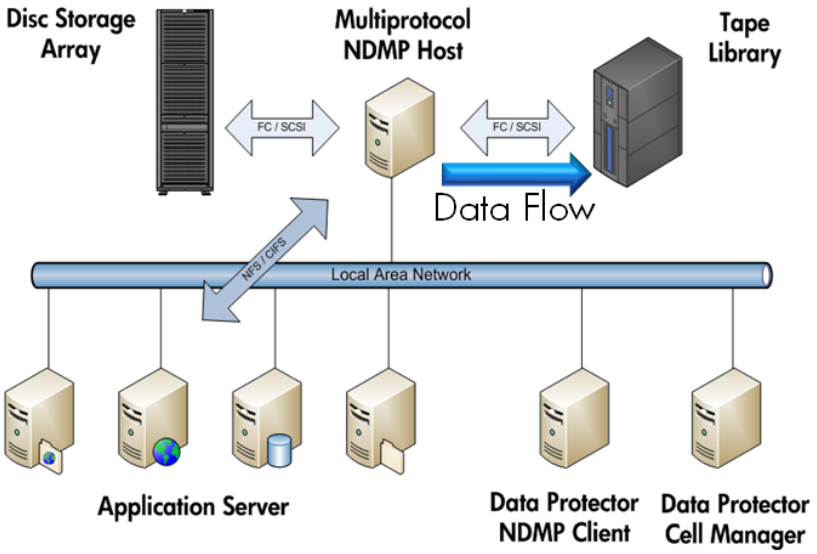
The full specification and additional NDMP information is published at http://www.ndmp.org/info/.

NDMP backup

In a typical environment, shown in Figure 24, the NDMP host and the Data Protector client with the NDMP Media Agent installed (NDMP client) are connected to the LAN. However, data from the NDMP server disks does not flow through the LAN; it is backed up to a tape device connected to the NDMP server system. The NDMP client initiates, monitors, and controls data management, and the NDMP server executes these operations, having direct control over devices connected to it and over the backup and restore speed.

Figure 24: NDMP and Data Protector

- Client/server model
- Data Protector is the NDMP client to the NDMP-Host Data Mover (Server)
- Backup data flows without traversing the network



Performance

In business-critical environments, it is a key requirement to minimize the time needed for data recovery in the case of a corrupt database or a disk disaster. Therefore, understanding and planning backup performance is extremely important. However, optimizing the time required for the backup of a number of client systems and large databases that are all connected on different networks and different platforms is a challenging task.

The following sections give an overview of the most common backup performance factors. Because of the high number of variables, it is not possible to give distinct recommendations that fit all user requirements.

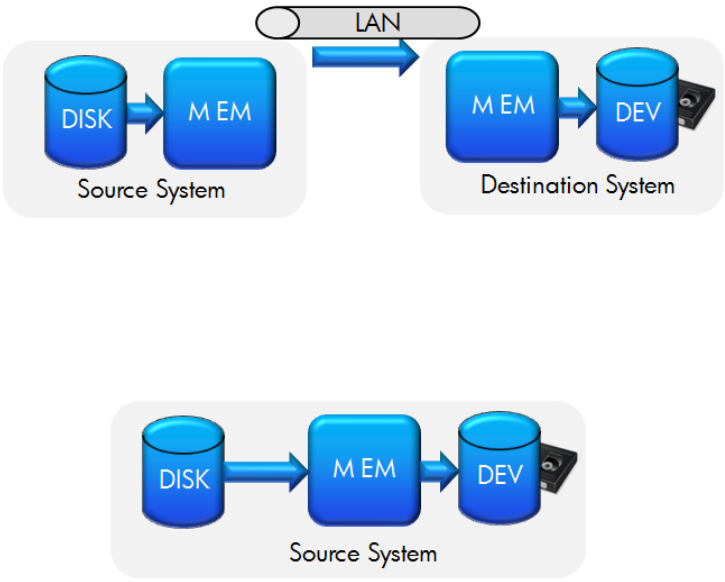
Network versus local backups

Sending data over a network introduces additional overhead, as the network becomes a component of performance consideration. Data Protector handles the data stream differently for network backup and local backup, as shown in figure 25:

Figure 25: Network versus local backup

– Network data stream flow
  1. Disk to memory of source system
  2. to network
  3. to memory of destination system
  4. to device

– Local data stream flow
  1. Disk to memory
  2. to device



To maximize performance, use local backup (data stream) configurations for high-volume data streams.

Devices

Device types and models impact performance because of the sustained speed at which devices can write data to a disk or tape (or read data from it).

Data transfer rates also depend on the use of hardware compression. The compression ratio that can be achieved depends on the nature of the data being backed up. In most cases, using high-speed devices with hardware compression improves performance. This is true, however, only if the devices stream.

At the start and at the end of a backup session, tape backup devices require some time for operations such as rewinding media and mounting or unmounting media.

Libraries offer additional advantages because of their fast and automated access to a large number of media. At backup time, new or reusable media must be loaded, and at restore time the media that contain the data to be restored need to be accessed quickly.

Data in disk-based devices is accessed faster than that in conventional devices, as there is no need to load and unload media. This reduces the amount of time spent for backup and restore. Additionally, disk-based devices enable the use of advanced backup strategies such as synthetic backup and disk staging, which also reduce the backup and restore time.

Computer systems

The speed of computer systems themselves directly impacts performance. The systems are loaded during backups by reading the disks, handling software compression, and so on.

The disk-read data rate and CPU usage are important performance criteria for the systems themselves, in addition to I/O performance and network types.

Advanced high-performance configuration

The Data Protector zero downtime backup solution provides a means of shortening the application downtime or backup mode time, and reduces the network overhead by using locally attached backup devices instead of network backup devices. The application downtime or backup mode time is limited to the time needed to create a replica of data, which is then backed up on a backup system to a locally attached device.

Using hardware in parallel

Using several data paths in parallel is a fundamental and efficient method of improving performance. This includes the network infrastructure. Parallelism boosts performance in the following situations:

- Several client systems can be backed up locally, that is, with disks and related devices connected on the same client system.
- Several client systems can be backed up over the network. Here the network traffic routing needs to be such that data paths do not overlap, otherwise performance is reduced.
- Several objects (disks) can be backed up to one or several (disk/tape) devices.
- Several dedicated network links between certain client systems can be used. For example, if system A has six objects (disks) to be backed up, and system B has three fast tape devices, consider using three dedicated network links between system A and system B.
- With Data Protector Load Balancing, Data Protector dynamically determines which object (disk) should be backed up to which device. Enable this feature, especially to back up a large number of filesystems in a dynamic environment.

Software compression

Software compression is performed by the client CPU when reading data from a disk. This reduces the data that is sent over the network, but it requires significant CPU resources from the client.

By default, software compression is disabled. Use software compression only for backups of many machines over a slow network, where data can be compressed before sending it over the network. If software compression is used, hardware compression should be disabled, since trying to compress data twice actually expands the data.

Hardware compression

Hardware compression is performed by a device that receives original data from a media server and writes it to media in the compressed mode. Hardware compression increases the speed at which a tape drive can receive data, because less data is written to the tape.

By default, hardware compression is enabled. On HP-UX systems, enable hardware compression by selecting a hardware compression device file. On Windows systems, enable hardware compression during device configuration. Use hardware compression with caution, because media written in compressed mode cannot be read using a device in uncompressed mode, and vice versa.

Full and incremental backups

A basic approach to improve performance is to reduce the amount of data to back up. Carefully plan your full and incremental backups. Consider that you may not need to perform all the full backups of all the client systems at the same time.

If you back up to disk, you can use advanced backup strategies such as synthetic backup and disk staging.

Disk image versus file system backups

It used to be more efficient to back up disk images (raw volumes, raw devices) rather than filesystems. This is still true in some cases, such as heavily-loaded systems or disks containing large numbers of small files. But disk image backups save all blocks, including unused blocks, which creates additional backup volume.

The general recommendation is to use filesystem backups if file-level restore is required.

Object distribution to media

The following are examples of object/media backup configurations provided by Data Protector:

- One object (disk) goes to one medium.

  The advantage is a known fixed relationship between an object and a medium on which the object resides. This can be of benefit for the restore process, since only one medium needs to be accessed.

  The disadvantage in a network backup configuration is the likely performance limitation due to the network, causing the device not to stream.

- Many objects go to a few media, each medium has data from several objects, one object goes to one device.

  The advantage here is the flexibility of data streams at backup time, helping to optimize performance, especially in a network configuration.

  The strategy is based on the assumption that the devices receive enough data to be able to stream, since each device receives data from several sources concurrently.

The disadvantage is that data (from other objects) has to be skipped during the restore of a single object. Additionally, there is no precise way of predicting which medium will receive data from which object.

Disk performance

All data that Data Protector backs up resides on disks in your systems, so the performance of the disks directly influences backup performance. A disk is essentially a sequential device, that is, you can read or write to it, but not both at the same time. Also, you can only read or write one stream of data at a time. Data Protector backs up filesystems sequentially, to reduce disk-head movements. It also restores files sequentially.

Disk fragmentation can also impair performance. Data on a disk is not kept in the logical order that you see when browsing the files and directories, but is fragmented in small blocks all over the physical disk. Therefore, to read or write a file, a disk head must move around the whole disk area. Note that this differs from one operating system to another.

SAN performance

The design of your SAN environment will affect the performance, efficiency, and reliability of your backup and recovery scheme. Inefficient SAN design can degrade the performance and efficiency of all members of the SAN.

For detailed SAN design considerations see the reference materials at http://www.hp.com/go/ebs.

Application performance

When you back up databases and applications, such as Oracle, SAP R/3, Sybase, and Informix Server, the performance of the backups also depends on the applications. Database online backups are provided so that backups can occur while the database application remains online. This helps to maximize database up time but may impact application performance. Data Protector integrates with all popular online database applications to optimize backup performance.

For more information on how Data Protector integrates with various applications and for tips on how to improve backup performance, see the appropriate *HP Data Protector Integration Guides*. Also see the documentation that comes with your online database application.

Creating your backup strategy

This section demonstrates backup strategies at a high level. The intention is more to provide a good overview with different examples than one detailed customer example including sizing.

Before creating your backup strategy, make sure that your requirements and assumptions are available and documented. The creation allows multiple options, based on the spread of assumptions, so the creation process will not result in one single solution. It is advisable to create not more than three options. The idea is to create a solution that fulfils the requirements, not one that uses all possible technologies and features.

Example 1 – VMware solution with Data Protector ZDB/IR

This example shows an enterprise backup environment with two or more VMware systems connected to an iSCSI storage array.

Requirements

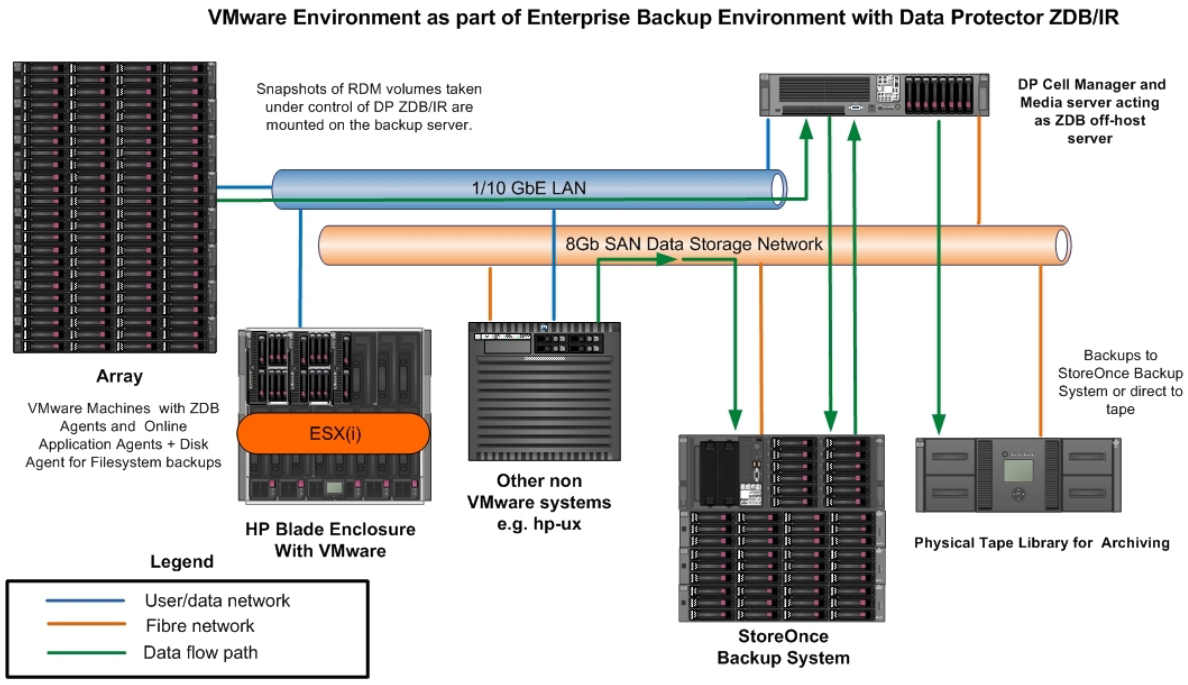
The customer has defined the following high-level requirements:

- The solution should be very flexible so that many backups and restores could run in parallel.
- The administration effort should be minimized.
- The customer does not want a high number of tape drives.
- The solution should enable fast file restores from the latest backup.
- Critical and large applications should be backed up outside the VMware environment (off-host).
- Critical applications should be restored very fast (instantly).
- Optionally, some important data should be protected for longer, and archived to tape.

Solution

The VMware solution in Figure 26 shows an HP StoreOnce Backup System and an HP tape library for archiving purposes. One Data Protector Cell Manager protects the complete VMware environment.

Figure 26: VMware solution with StoreOnce Backup System and MSL tape library



Data Protector provides the highest level of recovery point objective and recovery time objective (RPO/RTO) – also for virtualized mission-critical applications with zero impact to the virtual infrastructure. Together with HP Storage, Data Protector delivers fully-automated recovery of applications down to a specific point in time. With Data Protector ZDB/IR, data of critical and large applications is protected by Data Protector-managed array snapshots that are mounted on the Data Protector Cell Manager, backed up to the HP StoreOnce Backup System, and retained for Instant Recovery. Optionally, the data could be copied from the HP StoreOnce Backup System to an HP tape library.

Details

The hardware in the example consists of an HP StoreOnce D2D Backup System, which can be provided with either NAS or VTL backup targets. Any suitable model of StoreOnce Backup System could be used depending on the customer's storage capacity and performance requirements. The Data Protector Cell Manager is installed on an HP ProLiant server. The tape library could be an MSL2024 with one or two LTO5 tape drives.

Non-critical and smaller VMs are protected with the Virtual Environment Agent (VEAgent), which creates VM image backups. The VEAgent will communicate with the VMware vCenter Server and request a snapshot of the VM base disks to be backed up. Backup types can be full or incremental, including VMware Change Block Tracking. The snapshot is created in the storage array, and then the read-only image is presented on the backup server (Cell Manager) by the VEAgent. Backups of the VM images are sent to the Data Protector Logical Devices in the HP StoreOnce Backup System—VTL or NAS. Later, outside the backup window or perhaps at the end of month, Data Protector can perform an Object Copy operation from the StoreOnce to the tape library, giving the customer the ability of having a disk-based restore operation directly from the StoreOnce. Alternatively, if the backup sessions contained there have expired, the customer can perform a restore from the copy on the physical tape. Data Protector manages the retention period of the backup sessions.

Critical and large applications are protected by Data Protector Online and ZDB/IR Agents installed inside the virtual machine. Data Protector treats the virtual machine as a physical machine, which enables the highest level or protection. In VMware environments, all applications running as VMs can be protected as frequently as required, because all backups are processed by the storage array. Data Protector creates a snapshot (sometimes referred to as a replica or a clone) and moves that copy to a disk or tape for long-term storage. In addition, it can even be left on the disk for instant recovery. Data Protector Instant Recovery has the ability to recover snapshots to any point in time—down the exact second specified by the backup administrators—all from a single console. Data Protector takes advantage of what is called "database roll-forward functionality."

Example 2 – Remote and branch office solution

This example demonstrates how small and very small remote offices can be locally protected but also replicated to a central data center. All remote offices are centrally protected without any local backup administrator.

Requirements

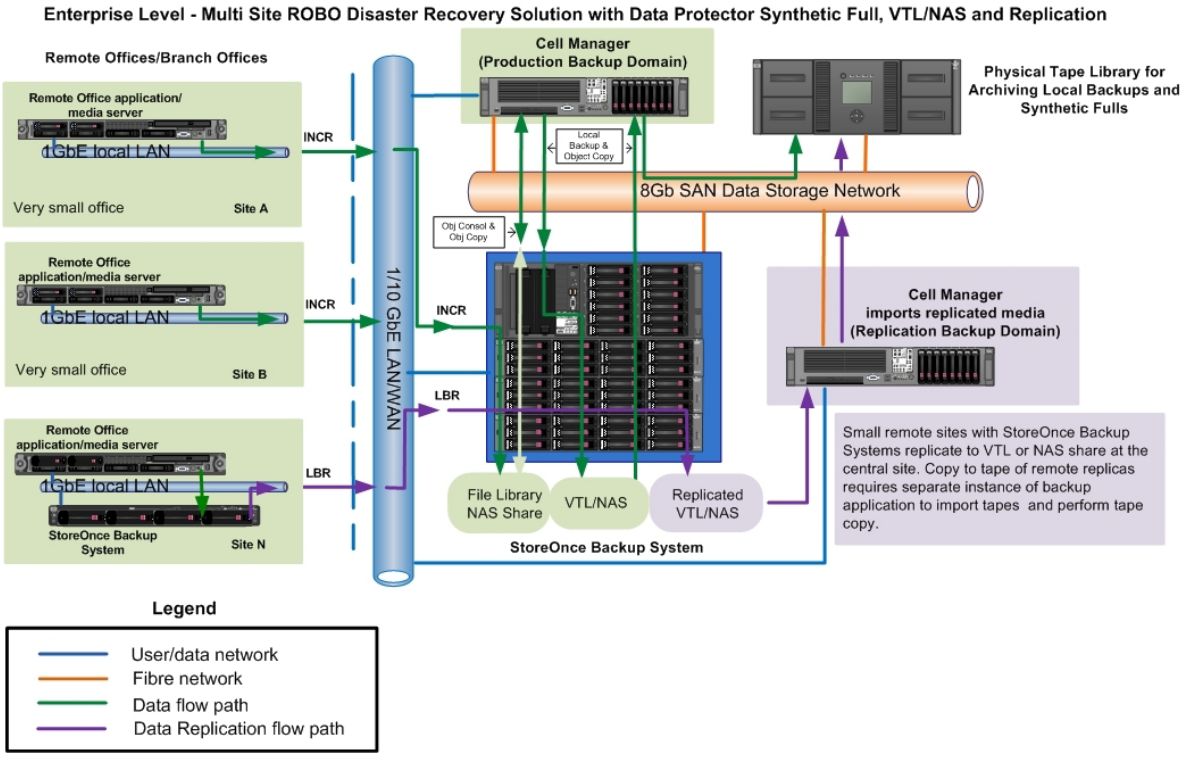
The customer requested a solution for the following scenario:

- Data of small and very small remote sites should be protected in the central data center.
- Remove physical tape, and limit human intervention for backups at small and very small remote sites.
- Consider long-term (multi-year) retention requirements for some backup data, which requires copying the backups to physical tape.
- Completely automate backup and disaster protection.
- Very small remote sites cannot justify the cost of a local appliance.

Solution

The enterprise-level solution in Figure 27 shows how all remote offices are protected by Data Protector.

Figure 27: Multi-site solution for remote and branch offices



Very small offices are protected by incremental-forever backups. The Data Protector Enhanced Incremental option saves data to a NAS share (Data Protector File Library) at the central site. The Data Protector Object Consolidation feature (Synthetic Full) will combine incrementals and a prior full backup to a new full backup set.

Other remote offices are protected by a local StoreOnce Backup System, which uses a NAS share or VTL emulation and replicates to another StoreOnce Backup System in the central data center.

Optionally, the Data Protector Object Copy feature could copy the replicated remote office data to a physical tape library for archiving purposes.

Details

Very small offices back up their data to a NAS share that is configured as a Data Protector File Library. Backups are performed with Enhanced Incremental Backup directly to the File Library. The first backup is a full backup and all following backups are incrementals. The Data Protector Object Consolidation feature creates a Synthetic Full

Backup from the prior full backup and the incremental backups, which is written back to the NAS share (File Library).

Other remote offices are saved locally and then replicated to the central data center. A separate Data Protector installation (Replication Backup Domain) is required in order to import the replicated data and copy it to tape. This migration to physical tape can happen outside the backup window.

Central site systems could be also protected by the StoreOnce Backup System, which would be configured as a VTL.

Note: The initial replication (seeding) of branch offices and also the first backup of remote offices could be very lengthy.

Recommendations for a successful implementation

While planning your solution with Data Protector, you can avoid some pitfalls to help save time and money:

1. **Learn about, and understand, Data Protector functionality**

   There are sometimes misunderstandings about Data Protector's features that lead to mistreatment and unexpected behavior.

   One example is that Data Protector organizes data within Backup Objects that cannot be compared with other backup software. Some backup administrators already know other products, start with Data Protector and then think that the architecture is the same. But there are some significant differences to consider. Data Protector's architecture was designed for easy handling, maximum flexibility and high performance. Data can be multi-streamed at client and backup-device level. Backup Objects can be filesystem or integration based. For filesystems, a Backup Objects is a backup unit that contains all items backed up from one disk volume (logical disk or mount point). The backed-up items can be any number of files and directories, or the entire disk or mount point. For integrations, a backup object represents the main components of a database or application, such as an entire database or sets of (archived) transaction logs.

2. **Leverage Data Protector functionality**

   Data Protector provides very powerful features for backup and copy. Multiplexing is one example. Multiple Disk Agents read data from the disk in parallel, and send the data to multiple Media Agents. This applies also to Object Copy, which is extremely flexible.

   Another example is synthetic backup. Synthetic backup is an advanced backup solution that eliminates the need to run regular full backups and puts a smaller load on backup clients. After an initial full backup, only incremental backups are run, which are subsequently merged with the full backup into a new, synthetic full backup. This can be repeated indefinitely, with no need to run a full backup again. Virtual full backup is an extended and more efficient version of synthetic backup. This solution uses pointers to consolidate data rather than copy the data. As a result, the consolidation takes less time and avoids unnecessary duplication of data.

3. **Consider performance requirements**

   Performance depends on many factors that must be carefully considered. Backup data streams through many components and devices, such as the CPU, LAN and tape device. One or several of them will be the limiting factor. It is not sufficient to base size on theoretical values, such the write performance of a tape device. Another issue is failing to consider restore times, or forgetting to ask business managers. Restores can take a very long time in case of busy components, which can occur particularly during business hours. The write performance of disk devices can also be slower if they are configured in RAID5 mode. This results in significantly different backup and restore times. Finally, it is very difficult to estimate restore times of deduplicated and staged (vaulted) data.

4. **Test in real production environments**

   In many cases, testing in productive environments is very useful to estimate backup and restore performance. Try to start testing during off-business hours. It need not be with Data Protector. You could also employ tools such as the HP StorageWorks Library and Tape Tools.

5. **Schedule and monitor housekeeping activities**

   No backup software is fully automated. Data Protector offers maintenance tools and jobs that reduce daily administrations tasks to a minimum, but it is important to understand that a new Data Protector installation

runs with pre-configured parameters that usually track more information than most businesses require. The result could be a fast-growing internal database, catalog and repository.

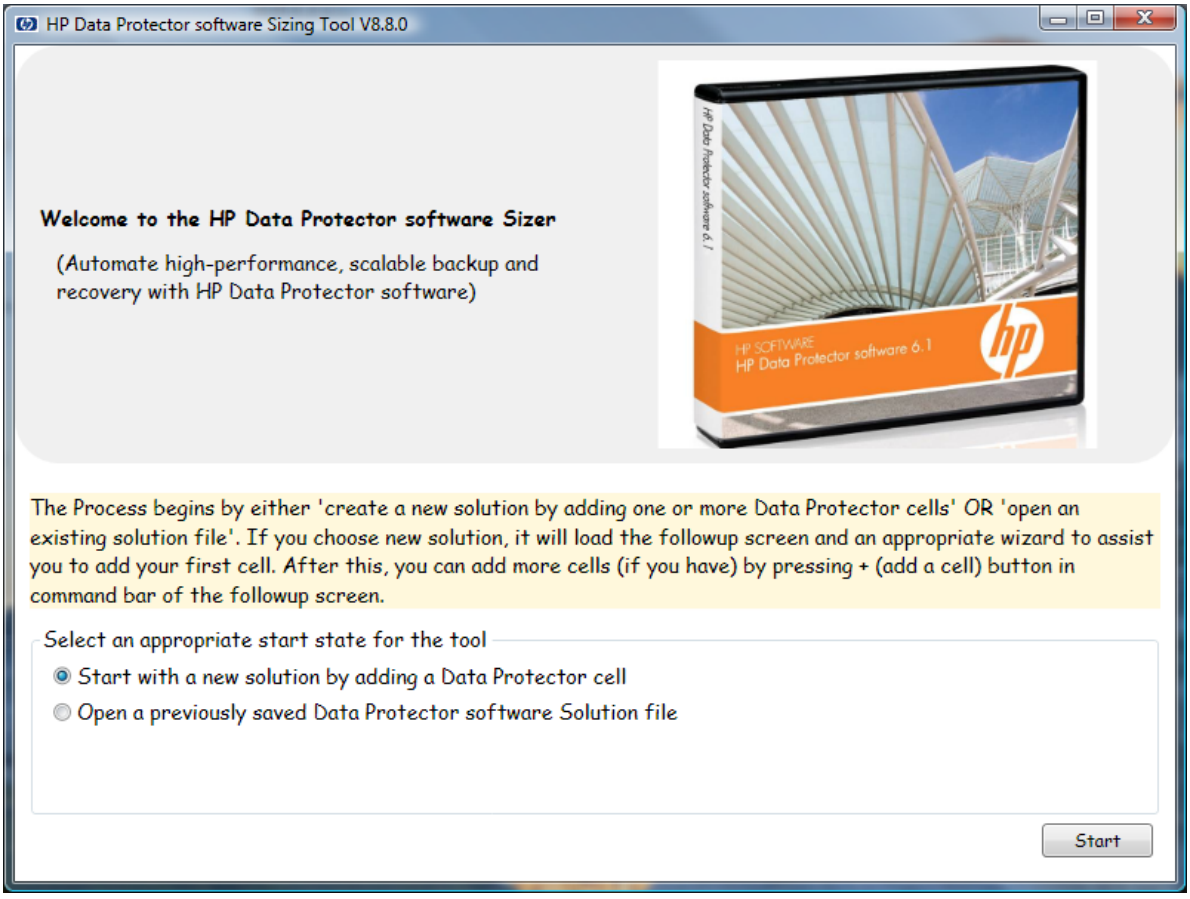
Appendix A – HP Data Protector software Sizing Tool

The HP Data Protector software Sizing Tool, shown in Figure 28, is part of the HP Storage Sizer. It is dedicated to help HP presales, partners and customers with sizing HP Data Protector software products and their components based on customer need. Starting with a high-level problem description, it helps customers with the end goal of quickly establishing what parts and services of Data Protector are appropriate for their needs.

Some of the design goals and features of the tool are:

- Easy to use user interface
- Guides users through the thought processes and sizing criteria involved in selecting software components
- Dedicated wizard-driven approach for each Data Protector cell
- Instant feedback as you start building your Data Protector cells/requirements incrementally
- Reporting view grouped by Data Protector cells
- Bill of materials with localized parts and pricing for many countries around the world, including the US
- Integrates well with HP Storage Sizer (also known as SWD Sizer)
- When invoked from HP Storage Sizer, combines the results of Data Protector Sizing Tool with the results of HP Storage Sizer so as to provide an integrated solution involving both storage and software
- Quick "size and quote" turnaround time

Figure 28: HP Data Protector software Sizing Tool



The sizing tool is downloadable from http://www.hp.com/go/storageworks/sizer.

Glossary of key terms

Application programming interface (API)

An application programming interface (API) is a particular set of rules ('code') and specifications that software programs can follow to communicate with each other. It serves as an interface between different software programs and facilitates their interaction, similar to the way the user interface facilitates interaction between humans and computers.

Cell

A set of systems that are under the control of a Cell Manager. The cell typically represents the systems on a site or in an organizational entity that are connected to the same LAN or SAN.

Cell Manager

The main system in the cell where the essential Data Protector software is installed, and from which all backup and restore activities are managed.

D2D2T (Disk to Disk to Tape)

D2D2T leverages the best features of both disk and tape storage to create a cost-effective, comprehensive data protection solution. Disk storage is used to automate and optimize daily backups, while providing instant access to data for fast restores.

Disk Agent (DA)

A Data Protector component needed on a client to back it up and restore it. The Disk Agent controls reading from and writing to a disk. During a backup session, the Disk Agent reads data from a disk and sends it to the Media Agent, which then moves it to the device. During a restore session the Disk Agent receives data from the Media Agent and writes it to the disk.

GUI (Graphical User Interface)

The Data Protector GUI is a cross-platform (HP-UX, Solaris, and Windows) graphical user interface, for easy access to all configuration, administration, and operation tasks.

HBA (Host Bus Adapter)

Connects a host system (the computer) to other network and storage devices.

IDB (Internal Database)

The Data Protector IDB is an internal database, located on the Cell Manager, that keeps information regarding what data is backed up, on which media it resides, the result of backup, restore, copy, object consolidation, and media management sessions, and which devices and libraries are configured.

IR (Instant Recovery)

Instant recovery restores a backup copy of data, held on the array, to its original location on the array to facilitate high-speed recovery. Data Protector ZDB and instant recovery techniques utilize mirror and snapshot technologies of disk-based arrays.

LAN (Local Area Network)

Computer network covering a small geographic area, like a home, office, or group of buildings.

MA (Media Agent)

Data Protector process that controls reading from and writing to a device, which reads from or writes to a medium (typically a tape). During a backup session, a Media Agent receives data from the Disk Agent and sends it to the device for writing it to the medium. During a restore session, a Media Agent locates data on the backup medium and sends it to the Disk Agent. The Disk Agent then writes the data to the disk. A Media Agent also manages the robotics control of a library.

MoM (Manager of Managers)

Allows administrators to manage a large environment, also known as an enterprise backup environment, using multiple Data Protector cells centrally from a single point.

NAS (Network Attached Storage)

File-level computer data storage connected to a computer network providing data access to heterogeneous clients. NAS not only operates as a file server, but is specialized for this task either by its hardware, software, or configuration of those elements. NAS is often made as a computer appliance—a specialized computer built from the ground up for storing and serving files—rather than simply a general purpose computer that is used for the role.

NDMP (Network Data Management Protocol)

A protocol developed by the NetApp and Legato companies, for transporting data between NAS devices and backup devices. It removes the need for transporting the data through the backup server itself, thus enhancing speed and removing load from the backup server.

RAID (Redundant Array of Independent Disks)

Computer data storage schemes that can divide and replicate data among multiple hard disk drives in order to increase reliability, allow faster access, or both.

ROBO (Remote Office Branch Office)

Off-site office that connects to the organization's WLAN or LAN externally.

RPO (Recovery Point Objective)

A point in time (prior to the outage) to which systems and data must be restored.

RTO (Recovery Time Objective)

A period of time after an outage within which the systems and data must be restored to the predetermined RPO (Recovery Point Objective).

SAN (Storage Area Network)

An architecture for attaching remote computer storage devices such as disk array controllers and tape libraries to servers, in such a way that the devices appear as locally attached devices to the operating system.

SLA (Service Level Agreement)

Part of a service contract where the level of service is formally defined. In practice, the term SLA is sometimes used to refer to the contracted delivery time (of the service) or performance.

VADP (vStorage API for Data Protection)

The next generation of VMware's data protection framework introduced in vSphere 4.0 that enables backup products to perform centralized, efficient, off-host LAN-free backup of vSphere virtual machines.

VEAgent (Virtualization Environment Integration Agent)

The Data Protector agent for VMware and Hyper-V support.

VSS (Volume Shadow Copy Service)

Shadow Copy (Volume Snapshot Service or Volume Shadow Copy Service or VSS), is a technology included in Microsoft Windows that allows the taking of manual or automatic backup copies or snapshots of data, even if it has a lock, on a specific volume at a specific point in time over regular intervals.

VTL (Virtual Tape Library)

Data storage virtualization technology used typically for backup and recovery purposes. A VTL presents a storage component (usually hard disk storage) as tape libraries or tape drives for use with existing backup software.

WORM (Write Once, Read Many)

A data storage device where information, once written, cannot be modified. WORM devices are useful in archiving information, where users want the security of knowing it has not been modified since the initial write.

ZDB (Zero Downtime Backup)

Data Protector ZDB and instant recovery techniques utilize mirror and snapshot technologies of disk-based arrays. ZDB creates, at high speed, a copy of the data to be backed up and then performs backup operations on the copy, rather than on the original data.

For more information

- HP Data Protector software
  http://www.hp.com/go/dataprotector
- HP Data Storage
  http://www.hp.com/go/storage
- Deduplication solutions
  http://www.hp.com/go/storeonce
- Enterprise Backup Solution (EBS)
  http://www.hp.com/go/ebs
- HP Library and Tape Tools
  http://www.hp.com/support/tapetools
- HP Storage Sizer
  http://www.hp.com/go/storageworks/sizer
- HP Technical Documentation
  http://www.docs.hp.com

Call to action

To read more about Data Protector, visit www.hp.com/go/dataprotector

4AA0-0789ENW, Rev. 2, Created November 2011