HP Dynamic Deduplication—
achieving a 50:1 ratio

Table of contents
Introduction ......................................................................................................................................... 2
Data deduplication—the hottest topic in data protection ................................................................. 2
The benefits of data deduplication ..................................................................................................... 2
How does data deduplication work? .................................................................................................. 3
What is Dynamic Deduplication? ...................................................................................................... 4
What deduplication ratio can I expect to achieve? ......................................................................... 4
HP test results .................................................................................................................................. 5
Test configuration ............................................................................................................................. 5
Backup scenario ............................................................................................................................... 5
Results .............................................................................................................................................. 6
Measurement .................................................................................................................................... 6
Conclusion ......................................................................................................................................... 6
For more information ....................................................................................................................... 7
Introduction

In June 2008, HP introduced two additional products to the HP StorageWorks D2D Backup System Family, the 1U rack-mountable D2D2500 and the 2U rack-mountable D2D4000. Both new products feature Dynamic deduplication based on an HP-patented deduplication algorithm that packs up to 50 times more backup data into the same disk footprint.

This paper explains how Dynamic deduplication works and also establishes how a 50:1 deduplication ratio may be achieved.

Data deduplication—the hottest topic in data protection

In these days of rampant data growth, a technology that can increase the effective capacity of a disk-based backup system by a ratio of up to 50:1 is big news.

Data deduplication allows you to store up to 50 times more backup data into the same disk footprint—giving you a better chance of restoring your users’ lost data from exactly the point they need it. And because fast restores of lost files are the hallmark of disk-based backup, your users are getting what they want when they want it. In any organization today, time is money, so restoring data and getting people back to work as quickly as possible saves money... and can save your reputation too.

The benefits of data deduplication

Adding data deduplication to disk-based backup delivers three key benefits:

1. It provides a cost-effective way of retaining your backup data on disk for longer—months instead of weeks—making file restores fast and easy from multiple recovery points. By extending data retention periods on disk, your backup data is easily accessible for longer periods of time before archiving to tape. In this way, lost or corrupt files can be quickly and easily restored from backups taken over a longer time span. This in turn improves service levels, reduces the disruption to your business, and allows users to return to work more quickly.

2. Efficient use of disk-space effectively reduces the cost-per-gigabyte of storage and can postpone the need to purchase more disk capacity. For data center managers who worry about the floor space and energy consumption required by storage hardware as data growth continues to accelerate, the HP disk-based backup systems with deduplication also reduce the space and power needs for a given volume of data.

3. Ultimately, data deduplication makes the replication of backup data over low-bandwidth WAN links viable (providing off-site protection for backup data) as only the changed data is sent across the connection to a second device (either a second identical device or one that comes from the same product family). Note that the HP StorageWorks D2D Backup Systems will not support data replication at introduction; however, this capability is under development for introduction later in 2008.
How does data deduplication work?

Essentially, data deduplication is the ability of a software application or an appliance to compare blocks of data being written to disk with data blocks that currently reside on that disk. Deduplication technology analyzes the incoming backup data and stores each unique block of data only once, using pointers for additional instances of the same data. When duplicate data is found, a pointer is established to the original set of data as opposed to actually storing the duplicate blocks—removing or “deduplicating” the redundant blocks from the volume.

A key point is that the deduplication is being done at the block level and not at the file level. In fact, beware of products that only dedupe at the file level (sometimes called “single instancing”).

Figure 1: How deduplication works

Say, for example, you are backing up a very large database that changes throughout the day. With a typical backup application, you have to back up, and more importantly store, the entire database with each backup. An incremental backup will not help you here. But with block-level deduplication, you can back up the same database to the device on two successive nights and, due to the ability to identify redundant blocks, only the blocks that have changed will be stored. Any redundant data will be stored as pointers instead.
What is Dynamic Deduplication?

HP Dynamic Deduplication uses an “inline” deduplication technique; that is, it deduplicates during the backup process, as opposed to post-processing techniques that deduplicate the data after the backup has been done. Inline deduplication provides a low-cost solution that does not require a lot of extra disk space to function.

The other advantage of Dynamic Deduplication is that it is application agnostic, so it is easy to implement in the majority of environments. This is due to the fact that it employs a hash-based process, rather than the object-oriented process used by Accelerated and other forms of deduplication. This hashing process has the advantage of working with any backup software.

Finally, Dynamic Deduplication is built into the latest HP D2D Backup Systems. It does not require a separate license.

What deduplication ratio can I expect to achieve?

There are a lot of numbers being quoted for deduplication ratios delivered by products in the industry, but what do they mean? Deduplication lets you store more for less, but how much space will you really save?

As with data compression, the actual data deduplication ratio you can achieve will vary and depends on:

1. The type of data being stored: The amount of duplicate data contained within each backup
2. Your backup policy: The more frequently you backup, the more you have to gain from data deduplication
3. The rate of data change: The less the data changes each day the higher the deduplication ratio

Deduplication ratios over one year of backing up a typical mix of business data could range from 10:1 up to 50:1, where a regime of daily full backups would produce a deduplication ratio in the high end of this range.

Another factor to consider is how data deduplication is measured by different manufacturers. In some cases, the ratios quoted do not consider traditional compression for the original data size. In other cases, the ratios are quoted for the deduplication achieved for the last backup (instead of considering the aggregate data stored).

The various dedup technologies in the industry (including technology provided by HP) would all deliver relatively similar ratios. What does affect the ratios quoted by vendors are the assumptions used for the above three points— backup policy, change rate, and how you measure.

To calculate deduplication ratios, HP uses the amount of data sent for backup as opposed to the amount actually stored, and takes both deduplication and data compression (at 2:1) into account. Note that some vendors may choose to base deduplication ratios from the point of a second backup and so discount the initial capacity used by the first backup to disk.
HP test results

The following test results were achieved in an independent test when measuring the deduplication ratio using Dynamic Deduplication with the HP D2D2500 and D2D4000 Backup Systems.

Test configuration

- Single backup server with
  - 2x dual-core AMD 2.4 GHz CPUs
  - 16 GB RAM
  - 3x 146 GB SAS HDDs in RAID0
  - Windows Server 2003 Enterprise Edition
- Connected through a dedicated LAN to a single iSCSI HP StorageWorks D2D4009i Backup System fully loaded with 12 disks
- Backup application used was Symantec Backup Exec 12d
- 4 GB base dataset to reflect typical business data
- The data incurred a daily change rate of approximately 0.4 %
- Initial backup data compression at 2.8:1

Backup scenario

To achieve a 50:1 deduplication ratio we performed nightly full backups using an extended-on-disk retention policy and appending to the previous night's backup, spanning virtual tapes when necessary without overwriting, deleting, or recycling any virtual media.

Data changed at a rate of 0.4 % per day. The ratio calculated was based on the amount of data sent for backup, as opposed to the amount actually stored, taking both deduplication and data compression into account.
Results

Figure 2: HP testbed deduplication results using Dynamic Deduplication

Achieving a 50:1 data deduplication ratio

Measurement

The deduplication ratio achieved was determined to be the amount of data sent by Backup Exec (the backup application) divided by the incremental raw disk space used to store that data.

Conclusion

Using representative business data sets in an independent testbed, a deduplication ratio of 50:1 was achieved after 35 full backup cycles.

Further independent testing is currently underway at client sites. The results of these tests will be published by HP as soon as they are available.
For more information

www.hp.com/go/d2d