Data deduplication and replication

Data deduplication is one of the most talked about new technologies in the storage industry today. And for good reason! Against a backdrop of rampant data growth, demanding security requirements, and urgent compliance needs, the IT industry has been developing disk-based data protection solutions that provide reliable automated backup, whilst ensuring fast access to data when things go wrong. Data deduplication technology builds on disk-based backup to let you store up to 50 times more data in the same disk footprint. This means users can keep more data at hand for longer, and it makes disk-based backup even more cost-effective!

As you would expect from a market leading company, HP offers a complete range of flexible, disk-based backup products featuring data deduplication—from entry-level models to enterprise-class virtual tape libraries.

But with so many deduplication solutions on the market today, it’s vital to understand how the technology works. HP uses target-based deduplication, where the processing is performed at the backup device or virtual library system, rather than using up valuable system processing resources.

So how does deduplication actually work? Let’s imagine we’re backing up a number of servers. Following an instruction from a backup application, each server sends a stream of data via the SAN or iSCSI connection to a virtual tape cartridge within the virtual tape library or autoloader. These backups are performed at regular intervals, typically on a daily basis.

Normally only a fraction of the total data held by a server changes between each backup as the majority of files and databases have only minor internal changes from one day to the next. Without deduplication an almost identical copy of the backup dataset will be stored each time a backup job is performed—and that takes up a lot of redundant disk space.

However, using the deduplication process, matching blocks of identical data already stored on the disk are identified, so that only one copy will be stored on the disk in its entirety. All subsequent occurrences of the same data block will be replaced by a pointer back to the original. As these pointers take up very little space, the more backups are performed—the greater the space gains.
Now, supposing that a data loss occurs and a file or database needs to be restored. In this instance, the data is simply reconstructed by the HP device back into its original undeduplicated state. It’s that simple!

With HP’s D2D systems and virtual library systems, we can also take advantage of traditional data compression which occurs before the data is stored on disk. And the larger HP virtual tape libraries use built-in hardware data compression for even faster performance.

But how much more deduplicated data can you expect to store on disk? The answer is that it depends. The deduplication ratios that can be achieved are really determined by the extent to which the data changes each day, and how long the data is retained on disk; however users can typically expect deduplication ratios of at least 20:1 for Microsoft Exchange and SQL Server backups. Meanwhile, independent tests commissioned by HP have revealed deduplication ratios of up to 50:1 over a three month period with typical rates of daily data change.

Being able to store more data on disk for longer is just one of the benefits of data deduplication. Another exciting outcome is the ability to automate your disaster recovery process and reduce cost by enabling site-to-site replication over low-bandwidth wide area networks. Because deduplication eliminates redundant data, replication transfers only the changed data between sites as opposed to the complete data set, dramatically reducing network costs. So for instance, as soon as the local backup at a remote site is complete, the central data centre can start to update the replica copy held on the virtual tapes at that end.

And in the unlikely event of a disaster wiping out the data at the remote site, new servers can be quickly rebuilt using the replicated data held in the data centre.

One size doesn’t fit all, though. That’s why HP offers you a choice of best-in-class data duplication solutions, with technology designed to fit individual business needs.

The HP D2D Backup System with integrated dynamic deduplication and replication radically changes the economics of data protection in smaller business environments and remote offices with a significantly lower price point than comparable products. The HP D2D allows an even wider range of businesses to benefit from deduplicated disk-based backup and low bandwidth data replication.

At the other end of the scale, there’s the HP StorageWorks Virtual Library System, incorporating accelerated deduplication technology. These remarkable appliances deliver data protection for the most demanding data centre environments—resulting in a truly scaleable solution with a capacity in excess of one petabyte.

So, there you have it. With data deduplication, HP’s family of virtual tape library solutions provide customers with a choice of disk-based solutions that pack even more backup data onto disk for reliable, cost-effective data protection. HP’s flexible deduplication solutions meet the varying needs of small and remote offices right up to the largest enterprise data centres.

No wonder deduplication and replication are 2 of the hottest topics in the storage industry today!