Quantum®

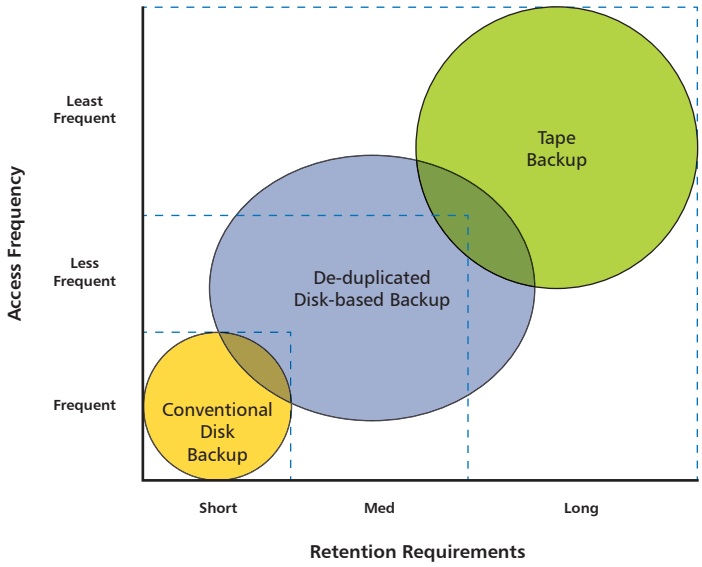
WHITE PAPER

D2D2T Backup Architectures and the Impact of Data De-duplication

CONTENTS

Introduction

The advent of data de-duplication, combined with disk-based backup systems, continues to change the landscape of backup and recovery environments. Conventional disk systems are well suited for short retention with frequent access, and tape are well suited for long-term retention with infrequent access. For many, data duplication technology fills the gap between the two, providing comparable performance to conventional disk but providing more cost-effective medium-term retention. De-duplication is also a key enabling technology for replication, making it practical for IT managers to link distributed sites and simplify media management across the enterprise.



This paper examines the impact of de-duplication in a disk-to-disk-to-tape (D2D2T) backup and recovery environment. In order to understand the impact of de-duplication, this paper provides background on the strengths and limitations of the current disk- and tape-based backup systems, as well as ideal applications of de-duplication to meet a wide range of customer needs.

This paper concludes by showing how an integrated, multi-tier backup system, leveraging de-duplication and remote replication, can provide a backup solution that combines a balance of short-term restore performance, long-term data retention, and overall cost effectiveness.

Strengths and Limitations of Tape Backup

Strengths of Tape

Tape-based backup continues to be a data protection mainstay for both small and large corporations. This is a testament to the fundamental utility and security of tape as a storage medium. It is also indicative of the ongoing advancement of tape technology in terms of capacity and cost, and continued improvements in the effectiveness of automated systems that integrate tape.

Tape drives, when provided with an uninterrupted high-speed data stream, are the performance front-runner in transferring data to a storage medium. The evolution of tape drive technology has also allowed tape to remain the leader in data density, providing the lowest cost per terabyte stored. A single LTO-4 drive can stream data natively at a rate of 120MB/second or 864GB/hour and can hold 800GB native with a compressed capability of 1.6TB per cartridge. A multi-drive library can easily backup and restore the multiple terabytes of data located in a typical data center with even the most aggressive SLA's or recovery time objectives.

Tape libraries, the systems that manage tapes and tape drives, are easily scaled and shared. In many situations, adding more performance or capacity to the library is as simple as adding more drives, more slots, and more media. Every major backup and recovery application has the ability to take advantage of tape libraries' scalability, as well as the nearly infinite amount of media that can be cycled through the system.

In addition to scalability, there are the benefits of portability and longevity. Since tapes are a removable medium, the ability to store them off-site in a secure location for years satisfies the majority of the long-term retention, disaster recovery, and compliance regulation requirements that companies are facing today. Using the built-in encryption available with LTO-4 and a capable key management system, securing data on tape is easier than ever – and does not impact the backup window.

Combining the high data density of tape with portability makes the transfer of very large data sets timely and cost effective. Transferring data over the network, even with compression and encryption, can consume valuable time and network resources. In many cases, shipping a set of encrypted tapes by trusted courier can be just as effective as transferring the data over the WAN and less susceptible to network bottleneck issues.

Recent generations of automated tape libraries have increased the overall effectiveness of tape-based protection by building in advanced reliability, availability, and serviceability functionality (RAS) to proactively monitor the critical library components and surrounding networks, and notify the administrative staff of any change in system status. Some tape system vendors have gone far beyond these features, incorporating additional features such as media integrity analysis, self-monitoring, auto-firmware leveling, and storage network diagnostics. Combining these advances provides the confidence necessary to maintain tape's position in the data protection strategy for a majority of the critical data in datacenters.

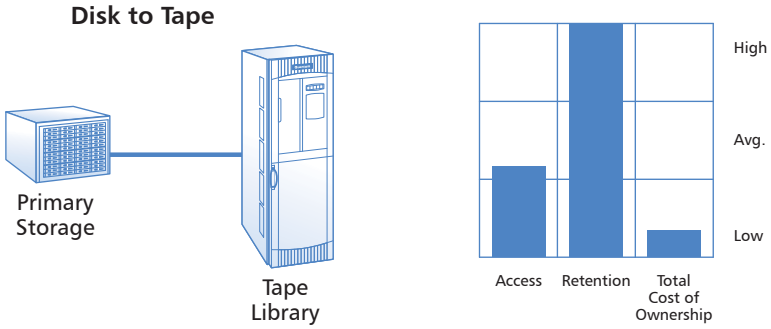
The greatest strength of tape is its combination of low cost and portability. While new disk technologies incorporating de-duplication have narrowed the cost and retention gap, tape is still by the far the low-cost leader for long-term retention and data density. Tape continues to have the lowest acquisition cost, the lowest cost for space, power, and cooling (making it the "greenest backup technology" available). In addition, with its media portability, it can transfer very large data sets in a fraction of the time and cost that it takes to transfer data between two disk systems on opposite sides of the country or the world.

Limitations of Tape

While tape is best suited for storing large, fast streaming data sets, it can be less effective at handling smaller, randomly accessed data sets. Backup times for multiple small data sets can be longer, depending on the tape append settings used by the backup application.

The linear write techniques that give tape the potential to stream data effectively also have a down side. When files to be restored are located at the middle or at the end of the tape, the tape must first be mounted, and all preceding data has to be bypassed to get to the required file. Meeting a recovery time objective (RTO) can be further slowed when the data to be recovered is on a tape that has already been shipped off to another site for storage.

In small offices or remote facilities, non-IT staff is typically tasked with managing tape backup. This can create challenges and hidden costs for those remote locations, especially when there is a need for regular maintenance or addressing a malfunction.



The Ideal Environment for Tape

Tape is ideally suited for environments where IT administration is consistently available, where the tape performance characteristics match the backup system to supply data, when data must be retained for extended periods, and when the lowest cost per unit of storage is the primary consideration.

Strengths and Limitations of Conventional Disk-based Backup

Strengths of Conventional Disk-based Backup

Unstructured data (home directories, email, individual desktop systems) is the fastest growing part of many data protection efforts, and these datasets are generally the ones that result in the majority of file-based data recovery operations. The ability to support random access reads and writes is, therefore, a key performance consideration for backup systems that protect unstructured data. Disk drives do not require a steady stream of data for optimal performance, and they can randomly access information stored on them. Since the transfer of unstructured data from a file server to a backup device is not always presented in a steady stream, disk is often able to handle this operation better than tape.

There are two general approaches to using a disk system as the target for backup software: using conventional RAID arrays or using specialized disk backup systems that emulate tape libraries (Virtual Tape Libraries—VTLs). Either approach can improve backup and restore performance, and many backup systems today use a layer of disk to provide rapid backup and restore for at least a portion of their backup tasks. Disk backup systems also take advantage of the redundancy built into contemporary RAID arrays to provide a high level of data availability.

Pointing backup jobs to a disk that is presented as disk, LUN, or network share is similar to those used for primary storage, is generally suited to smaller data protection environments. As data volumes grow, it becomes more complex to share and provision the backup capacity, and over time, fragmentation can slow down performance. As data volumes grow, presenting the disk resource as a VTL becomes more common. VTL presentations can provide better streaming performance than plain disk, and also provide more efficient scaling and sharing.

The most recognized strength of disk for backup is the fast and easy access to data. Since a disk system has inherent fast access characteristics, it minimizes delays during restores, and data stored on disk typically begins to be read within a few seconds. By comparison, data stored on tapes in a library can take several minutes to access, and if the data is stored off-site, the delay may be hours or days.

Limitations of Conventional Disk-based Backup

The higher performance and reliability of disk-based backup comes at a higher cost. While not nearly as expensive as primary disk systems, disk-based backup systems have a much higher operating cost and a significantly higher acquisition cost than tape. Data stored on disk is inherently vulnerable to site loss or damage, no matter how effective the RAID protection is. With tape systems, media can easily be moved to provide disaster recover (DR) protection, but with data on a disk backup system, an additional process is required to give the data that level of protection.

Disk-based backup systems excel at short- and medium-term storage. Even if the data never changes or grows, standard disk-based backup is not a cost effective medium for long-term retention. It is an inefficient use of very expensive resources to keep a disk system up and running year after year if the data isn't being accessed on a consistent basis.

Conventional disk-based backup is the least "green" technology in the data center, especially when compared to tape and de-duplicated disk solutions for long term retention. The following chart compares disk with tape and de-duplicated disk in terms of space, power, and cooling.

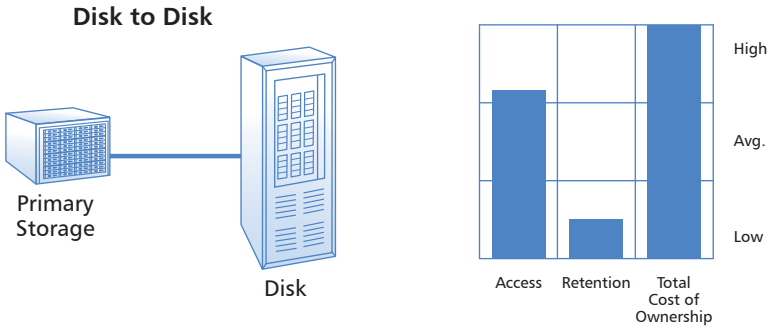
	Disk Array[1]	Tape Library[2]	Saving	Disk with data de-duplication[3]	Saving
Space	0.01m²/TB	0.00215m²/TB	78%	0.0003m²/TB	97%
Power	60 Watts/TB	.88 Watts/TB	99%	3.47 Watts/TB	94%
Cooling	204 BTU/TB/H	3.0 BTU/TB/H	99%	12.76 BTU/TB/H	94%

1. EMC CLARiiON CX20 with 120 750GB drives
2. Quantum Scalar i2000 with 300 slots, LTO-4, 2:1 compression
3. Quantum DXi5500, 10.8TB, 20:1 data de-duplication

In situations where disk is used as the only backup system (D2D) there can be capacity and retention problems. Unless more capacity is added, older backups must be expired in place so that the capacity can be reallocated for new backups. Forcing the deletion of data as a result of limited capacity may be seen as a benefit in some situations. Whether this strategy is truly beneficial in most cases is questionable, as almost every company has corporate or government rules that require it to retain data for a considerable number of years. Retaining data for years on disk is proven uneconomical, even with modern data reduction techniques.

Depending on the "flavor" of disk-based backup used (disk-as-disk, NAS, or VTL) other performance limitations may need to be considered. Some NAS deployments have exhibited network traffic bottlenecking, creating long backup windows and affecting the ability of users to access their files when the backup process runs long. Disk-as-disk implementations are subject to capacity management issues, requiring almost constant vigilance by administrative staff to ensure that disk volume provisioning matches all expected data growth. VTLs, with proprietary file systems, can create confusion around the location of data sets, particularly when replication is done independent of the data backup system.

The limitations with the greatest impact, and the one that has sparked the migration to disk-based de-duplication, is the tremendous cost of ownership for conventional disk, and the requirement for creating duplicate data sets to provide site loss protection. Even for data that does not require DR protection, depending on the size of the data set regularly backed up, the growth rate of the data, and the requirements for short- and long-term retention, conventional disk is a prohibitively expensive medium to use as the only backup system. Even with increasing disk system capacity, the continuing exponential growth of data makes disk alone a tough choice to justify in all but the most demanding high performance environments.



The Ideal Environment for Conventional Disk-based Backup

Conventional disk-based backup systems are ideally suited for high-performance data protection when the consistent streaming of data is not always possible, particularly when the source data is unstructured and transferred in a constant cycle of starts and stops. Disk backup is also ideal when there is a high frequency of restores, especially for small data sets or even individual files. In the overall strategy for a large or growing data management operation, conventional disk-based backup is best leveraged as a cache to shorten the backup window and provide additional time for data transfer to a longer retention medium such as tape. Disk backup systems are also ideally suited where disaster recovery is not a concern or where data does not need to be retained for months or years.

The Strengths and Limitations of De-duplicated Disk-based Backup

Defining Data De-duplication

As defined by the Storage Networking Industry Association (SNIA) – "Data De-duplication is accomplished by examining a data-set or I/O stream at the sub-file level, storing and/or sending only unique data. The definition of 'what is a duplicate' is predicated on the method used to evaluate, identify, track and avoid duplication. The de-duplication process includes updating tracking information, storing and/or sending data that is new and unique, and disregarding any data that is a duplicate."

Simply put, de-duplication eliminates redundant data at a sub-file level, using one of several systems to store only unique data. For backup, where most IT departments store highly redundant data sets over and over again, the value is extremely high.

The Strengths of De-duplication

De-duplication reduces disk capacity requirements by 90% or more for recurring backup operations, and it reduces the bandwidth required to replicate data by a similar factor by eliminating the need to transmit duplicate data elements. Customers benefit from storing more recovery points on the system and having the capability to restore data faster and more frequently. Since de-duplicated data requires less bandwidth and less time to transfer to another datacenter, remote replication is now a viable tool for disaster recovery—without de-duplication's bandwidth efficiency, few IT departments had the bandwidth or the time to replicate backups between locations.

With data de-duplication, it is becoming more common to see users storing weeks or months of daily backups on disk, replicate those datasets to off-site locations for short-term site-loss protection, and transfer only the data to tape that needs long term retention—often creating tapes only once a week or once a month. Most file restores are from local disk, but data is protected by having multiple copies available in different locations. As part of the overall tiered architecture, fewer tapes are shipped off-site for long-term retention and disaster recovery protection. With faster replication, the consolidation of data in one place is more easily accomplished, enabling the centralization of tape management. Instead of having multiple, small tape systems dispersed among a number of remote offices shipping tapes to a central datacenter, a de-duplicated disk system with remote replication capability can transfer datasets to a central system where copies can be written off to tape.

The most important strengths of de-duplicated disk are the opportunity to reduce the total cost of ownership by replacing racks full of conventional disk arrays, and to reduce the management of removable media in distributed environments by implementing a replication strategy. It is possible to achieve a measurable reduction in power, cooling, and datacenter footprint even with a modest de-duplication factor. In addition to lowering the Total Cost of Ownership (TCO) for disk backup operations, the remote replication features also lower operating expenses by enabling the elimination of tape systems and media at "remote offices".

The Limitations of De-duplication

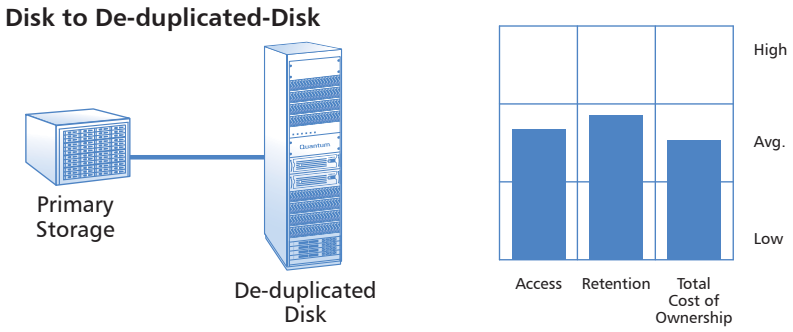
Data de-duplication is powerful technology but not every environment or type of data can benefit from it.

Environments with high data growth rates or high rates of change typically have too much unique data day to day to allow for any considerable reduction in the size of the backup data. For data that is almost entirely unique (encrypted data) or uncompressible (digital images), there is little benefit of using de-duplication. There may also be very little benefit for data that does not have to be retained over time, or where the data that is retained consists of images with significant amounts of unique data.

For example, four full backups created on four successive days will normally have a high degree of redundant data while four full backups created on the first day of each calendar quarter will normally have much less similarity.

Data de-duplication also inevitably introduces some amount of processing overhead which has an impact on performance, and different approaches to the technology change where the overhead is taken. Source-based de-duplication (where the technology is built into software running in the primary data environment), puts the overhead on production servers where it can slow down the backup job or the primary applications. Target based de-duplication (where an application sends data to a target system that de-duplicates it when it receives it) also has more than one way of handling the overhead. In-line systems process all the data during ingest, so the overhead has the potential to slow down the backup job.

Deferred, or post processing approaches, initially allow data to land on disk and then de-duplicate it outside the backup window—de-duplication won't slow down the backup but more disk is required. The adaptive method, which Quantum pioneered, de-duplicates during ingest but creates a disk buffer as well. It behaves like conventional in-line methods for most mid-range systems, but can adapt to faster ingest and avoid slowing down the backup so backup windows stay short. The newest generation of de-duplication systems—which the DXi7500 is the first—lets users pick different de-duplication methods for different backup tasks so users can get the right combination of performance and disk utilization for their unique mix of data.



The Ideal Environment for De-duplication

De-duplication systems are ideally suited for use in environments where the data being backed up has a reasonable amount of redundancy. If fact, the more redundancy in the data, the more effective the de-duplication system will be. A combination of highly redundant data with frequent, full backups maximizes the capacity savings and the restore performance.

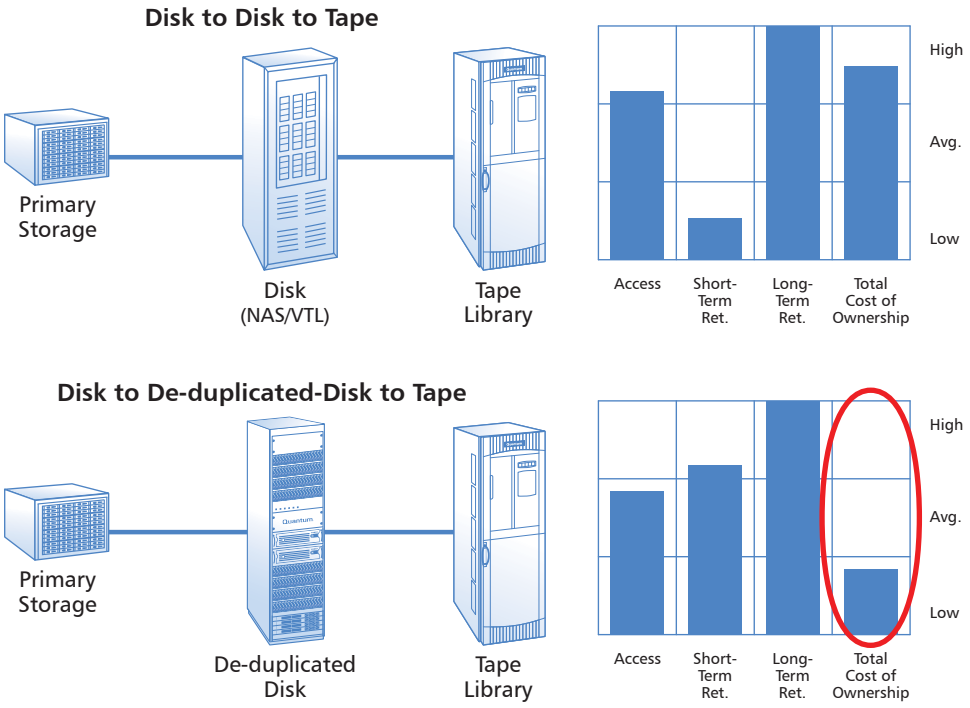
Examples of environments well suited for use with de-duplication include email systems, collaboration systems (with or without single instance storage capabilities), asset management systems, virtual servers (backing up the systems, their configurations, and related data), and home directories/file sharing systems.

Companies that want to reduce or eliminate tape systems in their regional or branch offices will be able to link their de-duplicated disk systems and securely replicate data between sites, automating their business continuance/disaster recovery processes. This in turn can reduce their administrative expenses and increase the overall reliability of their backup system.

D2D2T Architectures and the Impact of De-duplication

De-duplicated disk-based systems have changed the rules for D2D2T backup and recovery systems. De-duplicated disk backup systems, combined with tape and a capable data management application, enable customers to realize a broad range of benefits over and above the typical D2D2T systems:

- Meet more time sensitive service level agreements and their related RPOs and RTOs by retaining more backups on disk instead of tape
- Reduce operating costs (space, power, cooling) by operating fewer devices with more capacity
- Leverage faster replication to remove tape from remote sites, lowering operating and administrative expenses
- Centralize tape operations to increase security and limit risk of data loss
- Lower the overall TCO by reducing the number of disk systems required for backup as well as reducing the number of tapes created, managed, copied, shipped, and stored off-site



De-duplicated disk systems provide comparable performance and the same reliability of conventional disk systems with a significant capacity savings (90%+). This considerable improvement in capacity utilization allows the retention of more datasets for a longer period, providing greater restore performance to meet increasingly demanding RPOs and RTOs.

De-duplication makes remote replication faster and easier. With a de-duplicated backup, only the unique data is transferred to the centralized storage system. In the same way that capacity is optimized locally, data transfer is also optimized. With efficient replication, it is now possible to replace tape libraries in remote offices with de-duplicated disk systems and replicate data to a single data center or disaster recovery site where all tape creation tasks can be centrally administered.

De-duplication Enables More Powerful Remote Replication

Initial implementations of remote replication relied on bandwidth saving techniques, typically some form of in-line compression, to squeeze more data over existing networks and avoid expensive upgrades. Compression techniques are no longer able to keep pace with explosive data growth. A better solution for effectively transferring large datasets was established as one of the benefits of de-duplication.

By enabling the transfer of only new, unique information, de-duplication significantly reduces the necessary bandwidth (typically by 90% or more) for replicating large data sets. The extension of this capability is multi-site de-duplication and replication. Multi-site de-duplication and replication limits the amount of data transferred from one site to another and reduces data transfer from any other sites that have the same duplicate data on their systems. Essentially, multi-site replication scales the benefits of de-duplication from a single site to all of the connected sites.

The Impact of De-duplication on the Total Cost of Ownership (TCO)

Adding de-duplication to a typical D2D2T architecture can reduce the overall cost of a D2D2T backup system by nearly 400%, according to a recent paper published by The Clipper Group[2][3]. The paper calculates the cost ratio to store a terabyte of data on conventional SATA disk compared to LTO-4 tape at 23:1 without using de-duplication. When de-duplication is introduced, the cost savings of 90% disk capacity savings have a significant impact. The scenario is based on a multi-tier data protection solution, using disk-to-disk as a backup cache between either an automated tape library or conventional disk systems used for long-term retention.

De-duplicated disk backup systems can significantly alter the short-term retention and TCO for a tiered data protection solution. Even at a conservative de-duplication factor of 20x (The Clipper Group states "claims of effective de-duplication range from 10:1 to 100:1") it is conceivable that twenty racks of conventional disk-based backup systems can be replaced by just one rack of de-duplicated disk. This 20x increase in effective capacity dramatically reduces the power, cooling, and floor space requirements, all key factors in determining the TCO for disk-based backup systems.

In addition to the replacement of conventional disk, the use of de-duplication also provides the opportunity to alter the use of tape backup. By using de-duplication and replication from remote offices, datasets are protected as well as if they were written to tape and stored off-site. With this comparable protection in place, tape creation cycles can be altered – weekly or monthly full backups can be used instead of daily full/incremental backups. Depending on the environment, media usage can be curtailed at some level, allowing tape operations to focus more on longer-term data protection.

2 "Disk and Tape Square Off Again – Tape Remains King of the Hill with LTO-4", David Reine and Mike Kahn, www.clipper.com
3 While the analysis presented is this paper is very practical and well done [sounds like a casual opinion], the impact of de-duplication in the presented scenario does not take into consideration the reduction in media costs, as well as any reduction in shipping or long-term storage. Depending on the scale of the backup operation, the reduction of these factors would improve the cost savings even further.

Summary

No one technology by itself solves all the challenges for backup and recovery at every point in the data lifecycle. Different access and retention requirements, as well as cost considerations, can usually be satisfied with a multi-tier storage solution. By leveraging the strengths and benefits of a number of different integrated technologies, it is possible to fulfill these different requirements while lowering the overall TCO.

D2D2T backup systems have proven their ability to provide the necessary combination of speed and reliability along with short- and long-term retention. With the addition of de-duplicated disk-based backup systems, the benefits of D2D2T have been enhanced. More effective disk capacity provides better short-term to medium-term retention, while reducing the amount of tape needed for long-term retention, by making remote replication a viable option for a distributed enterprise.

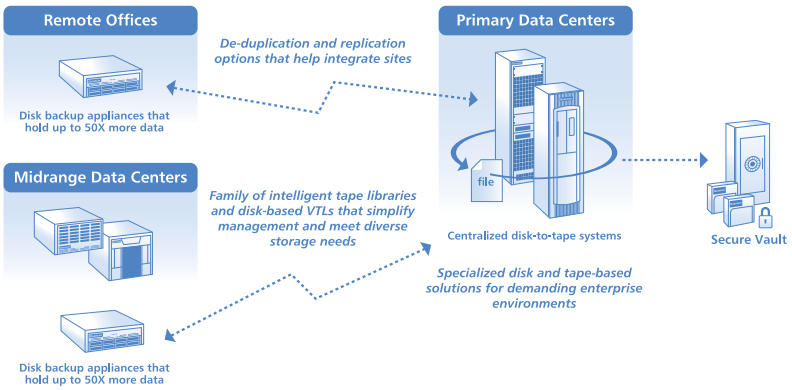
Achieving a balance between performance and cost is a matter of implementing each technology appropriately in the data protection lifecycle. De-duplication is a key technology that is bridging the gap between conventional disk and tape, bringing better balance between accessibility and retention, while improving TCO.

Quantum's Multi-tier Solution for Data Protection

Quantum offers D2D2T backup solutions using high performance virtual tape libraries (DX Series), disk based backup solutions with patented data de-duplication and remote replication (DXi Series), and a full range of automated tape libraries (Scalar® Series). For additional security, the tape libraries now offer encryption of data to be stored off-site and include key management.

Quantum has recently made it easy to combine all of these products together in a single system by introducing a common management tool—Quantum Vision—that discovers, manages and monitors all of these systems from a common console. In addition, all the systems can be covered by a single service and support organization, which makes a common software system available for remote monitoring and diagnostics.

Even at an individual solution level, Quantum is offering an integrated approach to multi-technology systems. The DXi7500, for example, is an enterprise disk backup system that combines multiple technologies at a solution level. It can operate as a conventional disk backup system using a native virtual tape library or NAS interface to provide the very highest throughput, as well as a data de-duplication enabled disk backup and remote replication system.



The DXi7500 features policy-based de-duplication, which allows users to apply different de-duplication methods to different backup tasks providing an optimal balance of performance and capacity savings. It also features a direct tape creation capability that automatically migrates backup data stored on disk to tape for longer term retention—and the movement takes place in the background without impacting the user's media server or backup SAN. The DXi7500's tape creation capability provides full support for backup software packages designed to initiate and manage automated data migration, including Symantec's NetBackup 6.5 Direct-to-Tape feature.

When adding encryption during replication (a core DXi feature), encryption on tapes for offsite storage, and common management across disk, tape, midrange, and Enterprise systems, end users have a highly integrated solutions set that makes it easy to combine the right balance of performance and cost over the life of a user's backup and retention.

For contact and product information,
visit quantum.com or call 800-677-6268

Quantum®
Backup. Recovery. Archive. It's What We Do.

About Quantum

Quantum Corp. (NYSE:QTM) is the leading global storage company specializing in backup, recovery and archive. Combining focused expertise, customer-driven innovation, and platform independence, Quantum provides a comprehensive range of disk, tape, media and software solutions supported by a world-class sales and service organization. As a long-standing and trusted partner, the company works closely with a broad network of resellers, OEMs and other suppliers to meet customers' evolving data protection needs.

WP00118 v2 Sep 2008