

Enhance storage efficiency
by removing redundant data!

Data De-duplication

FOR

DUMMIES[®]

Quantum Special Edition

**Reduce disk
storage needs
and lower costs!**

**A Reference
for the
Rest of Us!**

FREE eTips at dummies.com[®]

Mark R. Coppock
Steve Whitner



Data
De-duplication
FOR
DUMMIES®

QUANTUM SPECIAL EDITION

**by Mark Coppock
and Steve Whitner**



Wiley Publishing, Inc.

Data De-duplication For Dummies®, Quantum Special Edition

Published by

Wiley Publishing, Inc.

111 River Street

Hoboken, NJ 07030-5774

www.wiley.com

Copyright © 2008 by Wiley Publishing, Inc., Indianapolis, Indiana

Published by Wiley Publishing, Inc., Indianapolis, Indiana

Published simultaneously in Canada

No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning or otherwise, except as permitted under Sections 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 646-8600. Requests to the Publisher for permission should be addressed to the Legal Department, Wiley Publishing, Inc., 10475 Crosspoint Blvd., Indianapolis, IN 46256, (317) 572-3447, fax (317) 572-4355, or online at <http://www.wiley.com/go/permissions>.

Trademarks: Wiley, the Wiley Publishing logo, For Dummies, the Dummies Man logo, A Reference for the Rest of Us!, The Dummies Way, Dummies Daily, The Fun and Easy Way, Dummies.com, and related trade dress are trademarks or registered trademarks of John Wiley & Sons, Inc. and/or its affiliates in the United States and other countries, and may not be used without written permission. All other trademarks are the property of their respective owners. Wiley Publishing, Inc., is not associated with any product or vendor mentioned in this book.

LIMIT OF LIABILITY/DISCLAIMER OF WARRANTY: THE PUBLISHER AND THE AUTHOR MAKE NO REPRESENTATIONS OR WARRANTIES WITH RESPECT TO THE ACCURACY OR COMPLETENESS OF THE CONTENTS OF THIS WORK AND SPECIFICALLY DISCLAIM ALL WARRANTIES, INCLUDING WITHOUT LIMITATION WARRANTIES OF FITNESS FOR A PARTICULAR PURPOSE. NO WARRANTY MAY BE CREATED OR EXTENDED BY SALES OR PROMOTIONAL MATERIALS. THE ADVICE AND STRATEGIES CONTAINED HEREIN MAY NOT BE SUITABLE FOR EVERY SITUATION. THIS WORK IS SOLD WITH THE UNDERSTANDING THAT THE PUBLISHER IS NOT ENGAGED IN RENDERING LEGAL, ACCOUNTING, OR OTHER PROFESSIONAL SERVICES. IF PROFESSIONAL ASSISTANCE IS REQUIRED, THE SERVICES OF A COMPETENT PROFESSIONAL PERSON SHOULD BE SOUGHT. NEITHER THE PUBLISHER NOR THE AUTHOR SHALL BE LIABLE FOR DAMAGES ARISING HEREFROM. THE FACT THAT AN ORGANIZATION OR WEBSITE IS REFERRED TO IN THIS WORK AS A CITATION AND/OR A POTENTIAL SOURCE OF FURTHER INFORMATION DOES NOT MEAN THAT THE AUTHOR OR THE PUBLISHER ENDORSES THE INFORMATION THE ORGANIZATION OR WEBSITE MAY PROVIDE OR RECOMMENDATIONS IT MAY MAKE. FURTHER, READERS SHOULD BE AWARE THAT INTERNET WEBSITES LISTED IN THIS WORK MAY HAVE CHANGED OR DISAPPEARED BETWEEN WHEN THIS WORK WAS WRITTEN AND WHEN IT IS READ.

For general information on our other products and services, please contact our Customer Care Department within the U.S. at 800-762-2974, outside the U.S. at 317-572-3993, or fax 317-572-4002.

For technical support, please visit www.wiley.com/techsupport.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books.

Pocket Edition ISBN: 978-0-470-26054-8

Manufactured in the United States of America



Contents at a Glance

<i>Introduction.....</i>	<i>1</i>
<i>Chapter 1: Data De-duplication: Why Less Is More.....</i>	<i>3</i>
<i>Chapter 2: Data De-duplication in Detail.....</i>	<i>9</i>
<i>Chapter 3: The Business Case for Data De-duplication.....</i>	<i>21</i>
<i>Chapter 4: Ten Frequently Asked Data De-duplication Questions (and Their Answers).....</i>	<i>31</i>
<i>Appendix: Quantum's Data De-duplication Product Line.....</i>	<i>37</i>

Publisher's Acknowledgments

We're proud of this book; please send us your comments through our online registration form located at www.dummies.com/register/.

Some of the people who helped bring this book to market include the following:

Acquisitions, Editorial, and Media Development

Project Editor: Linda Morris

Acquisitions Editor:
Kyle Looper

Copy Editor: Linda Morris

Editorial Manager: Jodi Jensen

Composition Services

Project Coordinator:
Kristie Rees

Layout and Graphics:
Shelley Norris, Erin Zeltner

Proofreader: Melanie Hoffman

Publishing and Editorial for Technology Dummies

Richard Swadley, Vice President and Executive Group Publisher

Andy Cummings, Vice President and Publisher

Mary Bednarek, Executive Acquisitions Director

Mary C. Corder, Editorial Director

Publishing for Consumer Dummies

Diane Graves Steele, Vice President and Publisher

Joyce Pepple, Acquisitions Director

Composition Services

Gerry Fahey, Vice President of Production Services

Debbie Stailey, Director of Composition Services

Introduction



Right now, duplicate data is stealing time and money from your organization. It could be a presentation sitting in hundreds of user's network folders or a group e-mail sitting in thousands of inboxes. This redundant data makes both storage and protection more costly, more time-consuming, and less efficient. Data de-duplication, used on Quantum's DXi-Series disk backup and replication appliances, dramatically reduces this redundant data and the costs associated with it.

Data De-duplication For Dummies, Quantum Special Edition, discusses the methods and rationale for reducing the amount of duplicate data maintained by your organization. This book is intended to provide you with the information you need to understand how data de-duplication can make a meaningful impact on your organization's data management.

How This Book Is Organized

This book is arranged to guide you from the basics of data de-duplication, through its details, and then to the business case for data de-duplication.

- ✔ **Chapter 1: Data De-duplication: Why Less Is More:** Provides an overview of data de-duplication, including why it's needed, the basics of how it works, and why it matters to your organization.

- ✓ **Chapter 2: Data De-duplication in Detail:** Gives a relatively technical description of how data de-duplication functions, how it can be optimized, its various architectures, and how it can help with replication.
- ✓ **Chapter 3: The Business Case for Data De-duplication:** Provides an overview of the business costs of duplicate data, how data de-duplication can be effectively applied to your current data management process, and how it can aid in backup and recovery.
- ✓ **Chapter 4: Ten Frequently Asked Data De-duplication Questions (and Their Answers):** This chapter lists, well, frequently asked questions and their answers.

Icons Used in This Book

Here are some of the helpful icons you'll see used in this book.



The Remember icon flags information that you should pay special attention to.



The Technical Stuff icon lets you know that the accompanying text explains some technical information in detail.



A Tip icon lets you know that some practical information that can really help you is on the way.



A Warning icon lets you know of a potential problem that can occur if you don't take care.

Chapter 1

Data De-duplication: Why Less Is More

.....

In This Chapter

- ▶ Where duplicate data comes from
 - ▶ Identifying duplicate data
 - ▶ Using data de-duplication to reduce storage needs
 - ▶ Why data de-duplication is needed
-

Maybe you've heard the cliché "Information is the lifeblood of an organization." But many clichés have truth behind them, and this is one such case. The organization that best manages its information is likely the most competitive.

Of course, the data that makes up an organization's information must also be well-managed and protected. As amount and types of data an organization must manage increases exponentially, this task becomes harder and harder. Complicating matters is the simple fact that so much data is redundant.

To operate most effectively, every organization needs to reduce its duplicate data, increase the efficiency of its storage and backup systems, and reduce the overall cost of storage. Data de-duplication is one method for doing just that.

Duplicate Data: Empty Calories for Storage and Backup Systems

Allowing duplicate data in your storage and backup systems is like eating whipped cream straight out of the bowl: You get plenty of calories, but no nutrition. Take it to an extreme, and you end up overweight and undernourished. In the IT world, that means buying lots more storage than you really need.

The tricky part is that it's not just you who determines how much duplicate data you have. All of your users and systems generate duplicate data, and the larger your organization and the more careful you are about backup, the bigger the impact is.

For example, say that a sales manager sends out a 10MB presentation via e-mail to 500 sales people and each person stores the file. The presentation now takes up 5 GB of your storage space. Okay, you can live with that, but look at the impact on your backup!

Because yours is a prudent organization, each user's network share is backed up nightly. So day after day, week after week, you are adding 5 GB of data each

day to your backup, and most of the data in those files consists of the same blocks repeated over and over again. Multiply this by untold numbers of other sources of duplicate data, and the impact on your storage and backup systems becomes clear. Your storage needs skyrocket and your costs explode.

Data De-duplication: Putting Your Data on a Diet

If you want to lose weight, you either reduce your calories or increase your exercise. The same is sort of true for your data, except you can't make your storage and backup systems run laps to slim down.

Instead, you need a way to identify duplicate data and then eliminate it. *Data de-duplication* technology provides just such a solution. Systems like Quantum's DXi products that use block-based de-duplication start by segmenting a dataset into variable-length blocks and looking for repeated blocks. When they find a block they've seen before, instead of storing it again, they store a pointer to the original. Reading the file is simple — the sequence of pointers makes sure all the blocks are accessed in the right order.

Compared to other storage reduction methods that look for repeated whole files (single-instance storage is an example), data de-duplication provides a more granular approach. That means that in most cases, it dramatically reduces the amount of storage space needed.

As an example, consider the sales deck that we all just saved. Imagine that everybody put their name on the title page. A single-instance system would identify all the files as unique and save all of them. A system with data de-duplication, however, can tell the difference between unique and duplicate blocks inside files and between files, and it's designed to save only one copy of the redundant data segments. That means that you use much less storage.

A brief history of data reduction

One of the earliest approaches to data reduction was data *compression*, which searches for repetitive strings of information within a single file. Different types of compression technologies exist for different types of files, but all share a common limitation: Each reduces duplicate data only within individual files.

Next came *single-instance storage*, which reduces the amount of storage by recognizing when files are repeated. Single-instance storage is used in backup systems, for example, where a full backup is made first, and then incremental backups are made of only changed and new files. The effectiveness of single-instance storage is limited because it saves multiple copies of files that may only have minor differences.

Data de-duplication is the newest technique for reducing data. Because it recognizes differences at a variable-length block basis *within* files and *between* files, data de-duplication is the most efficient data reduction technique yet developed and allows for the highest savings in storage costs.

Data de-duplication isn't a stand-alone technology — it can work with single-instance storage and conventional compression. Data de-duplication can thus be integrated into existing storage and backup systems to decrease storage requirements without making drastic changes to an organization's infrastructure.



Data de-duplication utilizes proven technology. Most data is already stored in non-contiguous blocks, even on a single disk system, with pointers to where each file's blocks reside. In Windows systems, the File Allocation Table (FAT) maps the pointers. Each time a file is accessed, the FAT is referenced to read blocks in the right sequence. Data de-duplication references identical blocks of data with multiple pointers, but it uses the same basic system for reading multi-block files.

Why Data De-duplication Matters

Increasing the data you can put on a given disk makes sense for an IT organization for lots of reasons. The obvious one is that it reduces direct costs. Although disk costs have dropped dramatically over the last decade, the increase in the amount of data being stored has more than eaten up that savings.

Just as important, however, is that data de-duplication also reduces network bandwidth needs — when you store less data, you have to move less data too. That

opens up new protection capabilities — replication of backup data, for example — which make management of data much easier.

Finally, there are major impacts on indirect costs — the amount of space required for storage, cooling requirements, and power use.

Chapter 2

Data De-duplication in Detail

.....

In This Chapter

- ▶ Understanding how data de-duplication works
 - ▶ Optimizing data de-duplication
 - ▶ Defining the data de-duplication architectures
-

Data de-duplication is really a simple concept with very smart technology behind it: You only store a block once. If it shows up again, you store a pointer to the first occurrence, which takes up less space than storing the whole thing again. When data de-duplication is put into systems that you can actually use, however, there are several options for implementation. And before you pick an approach to use or a model to plug in, you need to look at your particular data needs to see whether data de-duplication can help you. Factors to consider include the type of data, how much it changes and what you want to do with it. So let's look at how data de-duplication works.

Making the Most of the Building Blocks of Data

Basically, data de-duplication segments a stream of data into variable length blocks and writes those blocks to disk. Along the way, it creates a digital signature — like a fingerprint — for each data segment and an index of the signatures it has seen. The index, which can be recreated from the stored data segments, lets the system know when it's seeing a new block.

When data de-duplication software sees a duplicate block, it inserts a pointer to the original block in the dataset's metadata (the information that describes the dataset) rather than storing the block again. If the same block shows up more than once, multiple pointers to it are created. It's a slam dunk — pointers are smaller than blocks, so you need less disk space.

A word about words

There's no science academy that forces IT writers to standardize word use — that's a good thing. But it means that different companies use different terms. In this book, we use *data de-duplication* to mean a variable-length block approach to reducing data storage requirements — and that's the way most people use the term. But some companies use the same word to describe systems that look for duplicate data in other ways — like at a file level. If you hear the term and you're not sure how it's being used, ask.

It's pretty clear that data de-duplication technology works best when it sees sets of data with lots of repeated segments. For most people, that's a perfect description of backup. Whether you back up everything everyday (and lots of us do this), or everything once a week with incremental backups in between, backup jobs by their nature send the same pieces of data to a storage system over and over again. Until data de-duplication, there wasn't a good alternative to storing all the duplicates. Now there is.

Fixed-length blocks versus variable-length data segments

So why variable-length blocks? You have to think about the alternative. Remember, the trick is to find the differences between datasets that are made up mostly — but not completely — of the same segments. If segments are found by dividing a data stream into fixed-length blocks, then changing any single block would mean that all the downstream blocks will look different the next time the data set is transmitted. Bottom line, you won't find very many common segments.

So instead of fixed blocks, Quantum's de-duplication technology divides the data stream into variable-length data segments using a system that can find the same block boundaries in different locations and contexts. This block-creation process lets the boundaries "float" within the data stream so that changes in one part of the dataset have little or no impact on

the blocks in other parts of the dataset. Duplicate data segments can then be found at different locations inside a file, inside different files, inside files created by different applications, and inside files created at different times. Figure 2-1 shows fixed block data de-duplication.

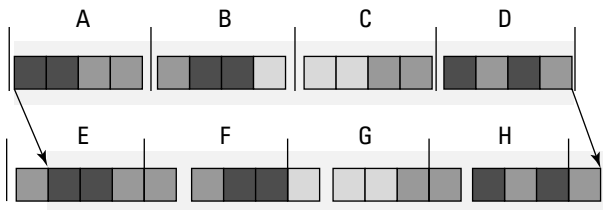


Figure 2-1: Fixed-length block data in data de-duplication.

The upper line shows the original blocks — the lower shows the blocks after making a single change to Block A (an insertion). The shaded sequence is identical in both lines, but all of the blocks have changed and no duplication is detected — there are eight unique blocks.

Data de-duplication utilizes variable-length blocks. In Figure 2-2, Block A changes when the new data is added (it is now E), but none of the other blocks are affected. Blocks B, C, and D are all identical to the same blocks in the first line. In all, we only have five unique blocks.

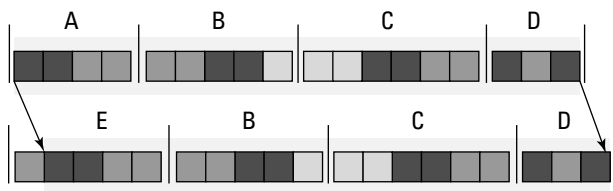


Figure 2-2: Variable-length block data in data de-duplication.

Effect of change in de-duplicated storage pools

When a dataset is processed for the first time by a data de-duplication system, the number of duplicate data segments varies depending on the nature of the data (both file type and content). The gain can range from negligible to 50% or more in storage efficiency.

But when multiple similar datasets — like a sequence of backup images from the same volume — are written to a common de-duplication pool, the benefit is very significant because each new write only increases the size of the total pool by the number of new data segments. In typical business data sets, it's common to see block-level differences between two backups of only 1% or 2%, although higher change rates are also frequently seen.

The number of new data segments in each new backup will depend a little on the data type, but mostly on the rate of change between backups. And total storage requirement also depends to a very great extent on your retention policies — the number of backup jobs and the length of time they are held on disk. The relationship between the amount of data sent to the de-duplication system and the disk capacity actually used to store it is referred to as the *de-duplication ratio*.

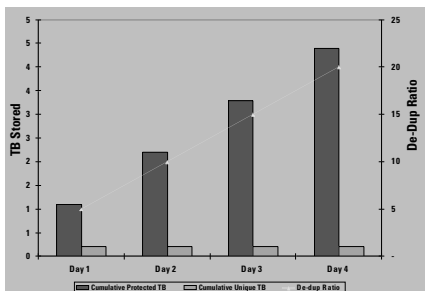
Figure 2-3 shows the formula used to derive the data de-duplication ratio. Figure 2-4 shows the ratio for four different backup datasets with different change rates (because compression also plays a part, this figure also shows different compression effects). These charts assume full backups, but de-duplication also works when incremental backups are included. As it turns out, though, the total amount of data stored in the de-duplication appliance may well be the same for either method because the storage pool only stores new blocks under either system. The data de-duplication *ratio* differs, though, because the amount of data sent to the system is much greater in a daily full model. So the storage advantage is greater for full backups even if the amount of data stored is the same.

$$\text{Data de-duplication ratio} = \frac{\text{Total data before reduction}}{\text{Total data after reduction}}$$

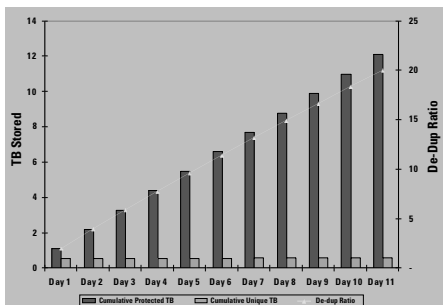
Figure 2-3: De-duplication ratio formula.

It makes sense that data de-duplication has the most powerful effect when it is used for backup data sets with low or modest change rates, but even for data

sets with high rates of change, the advantage can be very significant.

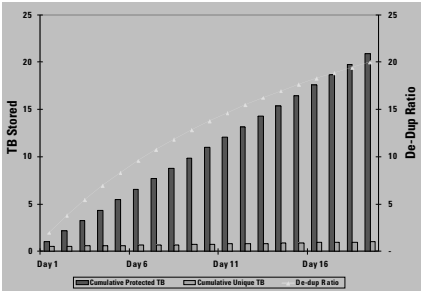


Backups for Data set 1
 Compressibility = 5:1
 Data change = 0%
 Events to reach 20:1 ratio = 4

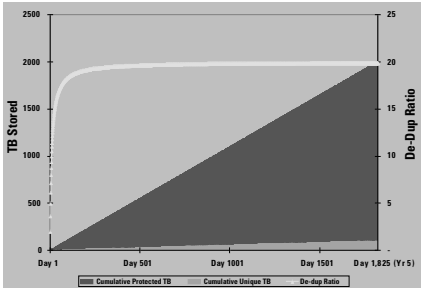


Backups for Data set 2
 Compressibility = 2:1
 Data change = 1%
 Events to reach 20:1 ratio = 11

Figure 2-4: Effects of data change on de-duplication ratios.



Backups for Data set 3
 Compressibility = 2:1
 Data change = 5%
 Events to reach 20:1 ratio = 19



Backups for Data set 4
 Compressibility = 2:1
 Data change = 10%
 Events to reach 20:1 ratio = 1825

Figure 2-4 (continued)

To help you select the right de-duplication appliance, Quantum uses a sizing calculator that models the growth of backup datasets based on the amount of data to be protected, the backup methodology, type of data, overall compressibility, rates of growth and change, and the length of time the data is to be retained. The sizing calculator helps you understand where data de-duplication has the most advantage and where more conventional disk or tape backup systems provide more appropriate functionality.



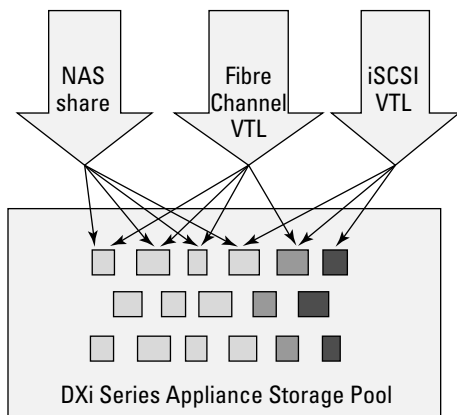
Contact your Quantum representative to participate in a de-duplication sizing exercise.

Sharing a Common Data De-duplication Pool

Many data de-duplication systems allow multiple streams of data from different servers and different applications to be sent into a common de-duplication pool (also called a *blockpool*) — that way, common blocks between different datasets can be de-duplicated. Quantum's DXi Series appliances are an example of such systems.

Each DXi Series system can present itself with different connection personalities at the same time — including any combination of NAS volumes (CIFS or NFS) and virtual tape libraries (VTLs, using either iSCSI or Fibre Channel protocol). Because all the presentations access a common blockpool, redundant blocks are eliminated across all the datasets written to the appliance. This means that a DXi-Series appliance

recognizes and de-duplicates the same data segments on a print and file server backed up via NAS and on an e-mail server backed up via VTL. Figure 2-5 demonstrates a sharing pool utilizing DXi-Series appliances.



Sharing storage pool in DXi Series appliances

All the datasets written to the DXi appliance share a common, de-duplicated storage pool irrespective of what presentation, interface, or application is used during ingest. One DXi Series appliance can support multiple presentations and interfaces simultaneously.

Figure 2-5: Sharing a de-duplication storage pool.

Data De-duplication Architectures

Data de-duplication, like compression or encryption, introduces some amount of overhead. So the choice of where and how de-duplication is carried out can affect backup performance. The most common approach today is to carry out de-duplication at the destination end of backup, but it can also occur at the source (that is, at the application server where the backup data is initially processed). When it takes place at the target end, the de-duplication process can be applied to data as it is ingested (*in-line processing*) or to data at rest on disk (*post processing*).

Wherever the data de-duplication is carried out, just as with compression or encryption, you get the fastest performance from purpose-built systems optimized for the process. If de-duplication is carried out using software agents running on general purpose servers, it's usually slower, you have to manage agents on all the servers, and de-duplication can compete with and slow down primary applications.

The data de-duplication approach with the highest performance and ease of implementation is generally carried out on specialized hardware systems at the destination end of backup data transmission. In this case, backup is faster and de-duplication can work with any backup software.

So of those target systems, which is better, in-line or post-processing de-duplication? The answer is . . . it depends. If you de-duplicate as you ingest (that's in-line), you need less disk space, keeping costs lower and using less space. This is the preferred choice for smaller systems — but it means that de-duplication overhead occurs during the backup window. For the highest performance initial backup, post-processing is your choice. This means you land all the data on disk first, and then do the de-duplication outside the backup window. The downside is that it takes more disk, but it's usually preferable for larger systems in very high performance environments.

As a note: Quantum de-duplicates at ingest on our midrange appliances — the DXi3500 and DXi5500. But on the Enterprise-scale DXi7500, they offer both. Users can do jobs using either approach. The only downside? It eliminates the argument about which is the better approach — and some people love to argue.

Chapter 3

The Business Case for Data De-duplication

.....

In This Chapter

- ▶ Looking at the business value of de-duplication
 - ▶ Finding out why applying the technology to replication and disaster recovery is key
 - ▶ Identifying the cost of storing duplicate data
 - ▶ Looking at the Quantum data de-duplication advantage
-

As with all IT investments, data de-duplication must make business sense to merit adoption. At one level, the value is pretty easy to establish. Adding disk to your backup strategy can provide faster backup and restore performance, as well as give you RAID levels of fault tolerance. But with conventional storage technology, the amount of disk people need for backup just costs too much. Data de-duplication solves that problem for many users because it lets them put 10 to 50 times more backup data on the same amount of disk.

Conventional disk backup has a second limitation that some users think is even more important — disaster

recovery (DR) protection. Can data de-duplication help there? Absolutely! The key is using the technology to power remote replication, and the outcome provides another compelling set of business advantages.

De-duplication to the Rescue: Replication and Disaster Recovery Protection

The minimum disaster recovery (DR) protection you need is to make backup data safe from site damage. After all, equipment and applications can be replaced, but digital assets may be irreplaceable. And no matter how many layers of redundancy a system has, when all copies of anything are stored on a single hardware system, they are vulnerable to fires, floods, or other site damage.

For most users, removable media provides site loss protection. And it's one of the big reasons that disk backup isn't used more: When backup data is on disk, it just sits there. You have to do something else to get DR protection. People talk about replicating backup data over networks, but almost nobody actually does it: Backup sets are too big and network bandwidth is too limited.

Data de-duplication changes all that — it finally makes remote replication of backup practical and smart. How does it work? Just like you only store the new blocks in each backup, you only have to replicate the new blocks. Suppose 1% of a 500GB backup has changed since the previous backup. That means you only have to move 5 GB of data to keep the two systems

synchronized — and you can move that data in the background over several hours. That means you can use a standard WAN to replicate backup sets.

For disaster recovery, that means you can have an off-site replica image of all your backup data every day, and you can reduce the amount of removable media you handle. That's especially nice when you have smaller sites that don't have IT staff. Less removable media can mean lower costs and less risk. Daily replication means better protection. It's a win-win situation.

How do you get them synched up in the first place? The first replication event may take longer, or you can co-locate devices and move data the first time over a faster network, or you can put backup data at the source site on tape and copy it locally onto the target system. After that first sync-up is finished, the replication only needs to move the new blocks.

What about tape? Do you still need it? Disk based deduplication and replication can reduce the amount of tape you use, but most IT departments combine the technologies, using tape for longer term retention. This approach makes sense for most users. If you want to keep data for six months or three years or seven years, tape provides the right economics and portability.

The best solution providers will help you get the right balance, and at least one of them — Quantum — lets you manage the disk and tape systems from a single management console, and it supports all your backup systems with the same service team.



The asynchronous replication method employed by Quantum in its DXi Series disk backup and replication solutions can give users extra bandwidth leverage. Before any

blocks are replicated to a target, the source system sends a list of blocks it wants to replicate. The target checks the list of proposed blocks against the blocks it already has, and then it tells the source what it needs to send. So if the same blocks exist in two different offices, they only have to be replicated to the target one time.

Figure 3-1 shows how the de-duplication process works on replication over a WAN.

Because many organizations use public data exchanges to supply WAN services between distributed sites, and because data transmitted between sites can take multiple paths from source to target, de-duplication appliances should offer encryption capabilities to ensure the security of data transmissions. In the case of DXi Series appliances, all replicated data — both metadata and actual blocks of data — is encrypted at the source level using SHA-AES 128-bit encryption and is decrypted at the target appliance.

Step 1:

Source sends a list of elements to replicate to the target. Target returns list of blocks not already stored there.

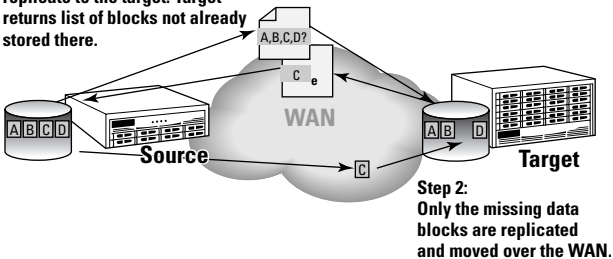


Figure 3-1: Verifying data segments prior to transmission.

Reducing the Overall Cost of Storing Data

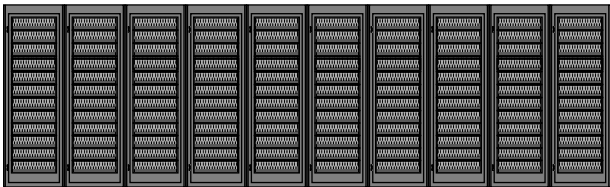
Storing redundant backup data brings with it a number of costs, from hard costs such as storage hardware to operational costs such as the labor to manage removable backup media and off-site storage and retrieval fees. Data de-duplication offers a number of opportunities for organizations to improve the effectiveness of their backup and to reduce overall data protection costs.

These include the opportunity to reduce hardware acquisition costs, but even more important for many IT organizations is the combination of all the costs that go into backup. They include on-going service costs, costs of removable media, the time spent managing backup at different locations, and the potential lost opportunity or liability costs if critical data becomes unavailable.

The situation is also made more complex by the fact that in the backup world, there are several kinds of technology and different situations often call for different combinations of them. If data is changing rapidly, for example, or only needs to be retained for a few days, the best option may be conventional disk backup. If it needs to be retained for longer periods — six months, a year, or more — traditional tape-based systems may make more sense. For many organizations, the need is likely to be different for different kinds of data.

The savings from combining disk-based backup, de-duplication, replication, and tape in an optimal way can provide very significant savings when users look at their total data protection costs. A recent analysis at a major software supplier showed how they could add

de-duplication and replication to their backup mix and save more than \$1,000,000 over a five-year period — reducing overall costs by about one-third. Where were the savings? In reduced media usage, lower power and cooling, and savings on license and service costs. The key was data de-duplication. If they tried the same approach using conventional disk technology, they would have *increased* costs — both because of higher acquisition expenses and much higher requirements for space, power, and cooling. (See Figure 3-2.)



Conventional Disk 1PB, 10 Racks

versus



Quantum's DXi5500 28:1 DeDup = 1PB, 20 U

Figure 3-2: Conventional disk technology versus Quantum's DXi5500.

The key to finding the best answer is looking clearly at *all* the alternatives. A supplier like Quantum that can provide and support all of the different options is likely to give users a wider range of solutions than a company that only offers one kind of technology.



Work with Quantum and the company's sizing calculator to help identify the right combination of technologies for the optimal backup solution both in the short term and the long term. See Chapter 2 for more on the sizing calculator.

Data De-duplication Also Works for Archiving

I've talked about the power of data de-duplication in the context of backup because that application includes so much redundant data. But data de-duplication can also have very significant benefits for archiving and nearline storage applications that are designed to handle very large volumes of data. By boosting the effective capacity of disk storage, data de-duplication can give these applications a practical way of increasing their use of disk-based resources cost effectively. Storage solutions that use Quantum's patented data de-duplication technology work effectively with standard archiving storage applications as well as with backup packages, and the company has integrated the technology into its own StorNext® data management software. Combining high-speed data sharing with cost effective content retention, StorNext helps customers consolidate storage

resources so that workflow operations run faster and the storage of digital business assets costs less. With StorNext, data sharing and retention are combined in a single solution that now also includes data de-duplication to provide even greater levels of value.

Looking at the Quantum Data De-duplication Advantage

The DXi-Series disk backup and replication systems use Quantum's data de-duplication technology to expand the amount of backup data users can retain on fast recovery RAID systems by 10 to 50 times. And they make automated replication of backup data over WANs a practical tool for DR protection. All DXi-Series systems share a common replication methodology, so users can connect distributed and midrange sites with Enterprise data centers. The result is a cost-effective way for IT departments to store more backup data on disk, to provide high-speed, reliable restores, to increase DR protection, to centralize backup operations, and to reduce media management costs.

The mid-range DXi Series appliances (the DXi3500 and DXi5500) offer eight different models with in-chassis scalability and industry-leading performance. The appliances are easy to install and use with all leading backup applications, and provide easy-to-use interface options including NAS, virtual library, or mixed presentations along with Fibre Channel and iSCSI connectivity.

The Enterprise DXi7500 offers true data center disk backup capabilities: scalability from 24 to 240 TB, performance up to 8 TB/hour, and full redundancy that uses active-active fail-over to eliminate any single point of failure throughout the system. The DXi7500 includes a direct tape creation capability to make it easy to centralize media management, and it is unique in giving users policy-based de-duplication — allowing them to chose a combination of in-line and post-processing de-duplication, depending on user needs.

All DXi solutions include Quantum's unique integrated software layer that includes the company's patented de-duplication techniques, high-performance file system technology, in-line compression, asynchronous replication, and built-in management, monitoring, alerting, and diagnostic tools.

DXi Series appliances are part of a comprehensive set of backup solutions from Quantum, the leading global specialist in backup, recovery, and archive. Whether the solution is disk with de-duplication, conventional disk, tape, or a combination of technologies, Quantum offers centralized management and unified service and support for all of your backup systems.

Chapter 4

Ten Frequently Asked Data De-duplication Questions (and Their Answers)

.....

In This Chapter

- ▶ Figuring out what data de-duplication really means
 - ▶ Discovering the advantages of data de-duplication
-

In this chapter, we answer the ten questions most often asked about data de-duplication.

What Does the Term “Data De-duplication” Really Mean?

There’s really no industry-standard definition yet, but there are some things that everyone agrees on. For example, everybody agrees that it’s a system for eliminating the need to store redundant data, and most people limit it to systems that look for duplicate data at a block level, not a file level. Imagine 20 copies of a presentation that have different title pages: To a file-level data reduction system, they look like 20 completely

different files. Block-level approaches see the commonality between them and use much less storage.

The most powerful data de-duplication uses a variable-length block approach. A product using this approach looks at a sequence of data, segments it into variable length blocks, and, when it sees a repeated block, stores a pointer to the original instead of storing the block again. Because the pointer takes up less space than the block, you save space. In backup, where the same blocks show up again and again, users can typically store 10 to 50 times more data than on conventional disk.

How Is Data De-duplication Applied to Replication?

Replication is the process of sending duplicate data from a source to a target. Typically, a relatively high performance network is required to replicate large amounts of backup data. But with de-duplication, the source system — the one sending data — looks for duplicate blocks in the replication stream. Blocks already transmitted to the target system don't need to be transmitted again. The system simply sends a pointer, which is much smaller than the block of data and requires much less bandwidth.

What Applications Does Data De-duplication Support?

When used for backup, data de-duplication supports all applications and all qualified backup packages. Certain file types — some rich media files, for example — don't see much advantage the first time they are sent

through de-duplication because the applications that wrote the files already eliminated redundancy. But if those files are backed up multiple times or backed up after small changes are made, de-duplication can create very powerful capacity advantages.

Is There Any Way to Tell How Much Improvement Data De-duplication Will Give Me?

Four primary variables affect how much improvement you will realize from data de-duplication:

- ✓ How much your data changes (that is, how many new blocks get introduced)
- ✓ How well your data compresses using conventional compression techniques
- ✓ How your backup methodology is designed (that is, full versus incremental or differential)
- ✓ How long you plan to retain the backup data

Quantum offers sizing calculators to estimate the effect that data de-duplication will have on your business. Pre-sales systems engineers can walk you through the process and show you what kind of benefit you will see.

What Are the Real Benefits of Data De-duplication?

There are two main benefits of data de-duplication. First, data de-duplication technology lets you keep more backup data on disk than with any conventional

disk backup system, which means you can restore more data faster. Second, it makes it practical to use standard WANs and replication for disaster recovery (DR) protection, which means users can provide DR protection while reducing the amount of removable media (that's tape) handling that they do.

What Is Variable-Block Length Data De-duplication?

It's easiest to think of the alternative to variable-length, which is fixed-length. If you divided a stream of data into fixed-length segments, every time something changed at one point, all the blocks downstream would also change. The system of variable-length blocks that Quantum uses allows some of the segments to stretch or shrink, while leaving downstream blocks unchanged. This increases the ability of the system to find duplicate data segments, so it saves significantly more space.

If the Data Is Divided into Blocks, Is It Safe?

The technology for using pointers to reference a sequence of data segments has been standard in the industry for decades: You use it every day, and it is safe. Whenever a large file is written to disk, it is stored in blocks on different disk sectors in an order determined by space availability. When you "read" a file, you are really reading pointers in the file's metadata that reference the various sectors in the right order. Block-based data de-duplication applies a similar kind of

technology, but it allows a single block to be referenced by multiple sets of metadata.

When Does Data De-duplication Occur During Backup?

There are really two choices.

You can send all your backup data to a backup target and perform de-duplication there, or you can perform the de-duplication on the host during backup. Both systems are available and both have advantages. If you de-duplicate on the host during backup, you send less data over your backup connection, but you have to manage software on all the protected hosts, backup slows down because de-duplication adds overhead, and you're using a general purpose server, which can slow down other applications. If you de-duplicate at the backup target, you send more data over the connection, but you can use any backup software, you only have to manage a single target, and the performance is normally a lot higher because the hardware system is specially built just for de-duplication.

Does Data De-duplication Support Tape?

Yes and no. Data de-duplication needs random access to data blocks for both writing and reading, so it must be implemented in a disk-based system. But tape can easily be written from a de-duplication data store, and, in fact, that is the typical practice. Most de-duplication

customers keep a few weeks or months of backup data on disk, and then use tape for longer-term storage. When you create a tape from de-duplicated data, the data is re-expanded so that it can be read directly in a tape drive and does not have to be written back to a disk system first. This is important because you want to be able to read those tapes directly in case of an emergency restore.

What Do Data De-duplication Solutions Cost?

Costs can vary a lot, but the 20:1 ratio is a pretty good rule of thumb starting point. Assuming an average de-duplication advantage of 20:1 — meaning that you can store 20 times more data than conventional disk, a number widely used in the industry — it's common to see list prices in the range of \$1/GB. A system that could retain 20TB of backup data would have a list price of around \$20,000. That's a lot lower than if you protected the same data using conventional disk. Note that this ratio is based on the manufacturer's suggested retail price: It's common to see discounts from both resellers and manufacturers.

Appendix

Quantum's Data De-duplication Product Line

In This Appendix

- ▶ Reviewing the Quantum DXi Series Disk Backup and Remote Replication Solutions
- ▶ Identifying the features and benefits of the DXi Series
- ▶ Examining the technical specifications of the DXi Series

Quantum Corp. 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, 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, original equipment manufacturers (OEMs), and other suppliers to meet customers' evolving data protection needs.

Quantum's DXi Series disk backup solutions leverage data de-duplication technology to increase disk capacities by as much as 10 to 50 times and replicate data

between sites over existing wide area networks (WANs). The DXi Series spans the widest range of backup capacity points in the industry, from 1.2TB to 240TB of raw disk capacity. All models share a common software layer, including de-duplication and remote replication, allowing IT departments to connect all their sites in a comprehensive data protection strategy that boosts backup performance, reduces or eliminates media handling, and centralizes disaster recovery operations.

The DXi3500 and DXi5500 are easy-to-use disk backup appliances for distributed offices and midrange data centers. The DXi7500 offers enterprise performance and capacity, along with direct tape creation, policy-based de-duplication, and high-availability architecture. All DXi solutions offer simultaneous NAS and VTL interfaces, and all are supported by Quantum, the leading global specialist in backup, recovery, and archive. Figure A-1 shows how DXi Series replication uses existing WANs for DR protection, linking backup data across sites and reducing or eliminating media handling.

Features and benefits of Quantum's DXi Series:

- ✓ Patented data de-duplication technology increases backup capacity of disk by 10 to 50 times
- ✓ Remote replication reduces bandwidth needed to move data, provides disaster recovery protection on existing WANs, and reduces or eliminates media handling
- ✓ Broad solution set — from 1.2TB to 240TB raw capacity — protects both distributed sites and large data centers
- ✓ Common software layer links all models and interfaces (NAS or VTL) in a comprehensive data protection strategy

- ✓ Easy-to-use appliances support distributed and midrange environments
- ✓ True enterprise capability — 8TB/hour performance, 240TB capacity, direct tape creation, and high availability architecture — creates a secure, centralized backup system
- ✓ Solutions work with all leading backup software, and can be added easily to existing environments

Quantum's Replication Technology

Users can transmit data from a single site or multiple sites to a central location over existing WANs for automated DR protection.

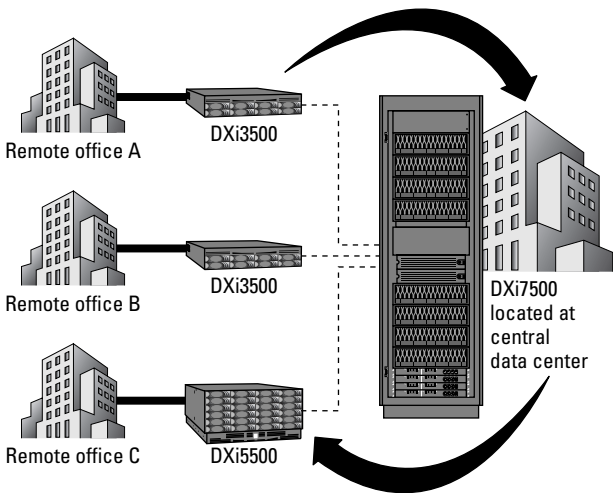


Figure A-1: DXi Series replication.

Tables A-1 through A-10 give the specifications for different DXi Series models.

Table A-1 DXi3500 Disk Backup Appliance Models

Usable capacity	1.2TB	1.8TB	2.8TB	4.2TB
Raw capacity	2TB	3TB	4TB	6TB
Retention capacity @ 20:1*	24TB	36TB	56TB	84TB
Retention capacity @ 50:1*	60TB	90TB	140TB	210TB

Table A-2 DXi5500 Disk Backup Appliance Models

Usable capacity	3.6TB	5.4TB	7.2TB	10.8TB
Raw capacity	6TB	9TB	12TB	18TB
Retention capacity @ 20:1*	72TB	108TB	144TB	216TB
Retention capacity @ 50:1*	180TB	270TB	360TB	540TB

* Capacities and amounts of data protected assume standard business data mix and extended on-disk retention. 20:1 capacity ratio assumes a weekly full and daily incremental backup model. 50:1 capacity ratio assumes daily full backups. Actual results will vary with data type, change rates, and backup methodologies. Smaller DXi Series appliances can be scaled in-chassis and on-site to specific larger units: Dxi3500 1.2TB units scale to 2.8TB units; DXi3500 1.8TB units scale to 4.2TB units; DXi5500 3.6TB units scale to 7.2TB units; DXi5500 5.4TB units scale to 10.8TB units. Data de-duplication is in-line.

Table A-3 DXi7500 Disk Backup Systems

Models	24TB to 240TB raw capacity
Policy based data de-duplication options	In-line: data is de-duplicated on ingest Post-processing: data is ingested to disk first, and de-duplicated in a separate process Both methodologies may be enabled for different data sets in the same DXi7500
Direct tape creation	Physical tape can be written in background over a dedicated Fibre Channel connection without using media server or backup SAN. Function maintains barcode integrity between virtual and physical tapes and is compatible with backup software direct-to-tape commands (for example, NetBackup 6.5).
High availability	Dual main system controllers (active-active) Dual RAID controllers (active-active) Redundant power Redundant cooling Hot swap drives, power supplies, fans

Table A-4 Interfaces — All Models

NAS backup target	NFS or CFS
Virtual library	Fibre Channel or iSCSI connectivity. Different partitions in same appliance can present different interfaces simultaneously.

Table A-5 Performance

DXi3500	Up to 290GB/hour
DXi5500	Up to 800GB/hour
DXi7500	Up to 8TB/hour

DXi3500, DXi5500, and DXi7500 models all offer support for remote replication. Replication is asynchronous, one-to-one or multiple-to-one configurations; partitions in the same unit act as replication source or target; units with partitions acting as replication targets can also support local backup.

Table A-6 Host to Appliance H/W Interface

DXi3500/DXi5500:	10/100/1000 BaseT Ethernet, 2Gb Fibre Channel
DXi7500	10/100/1000 BaseT Ethernet, 4Gb Fibre Channel

Table A-7**Power Input**

DXi3500	Power cord NEMA 5-15P to C13 3 connectors
DXi5500	Power Cord NEMA 5-15P to C13 4 connectors
Dxi7500	Power Cord NEMA L6-30P on system 2-8 connectors

Table A-8**Power Environment**

	<i>DXi3500</i>	<i>DXi5500</i>	<i>DXi7500</i>
Input voltages	100 to 240 VAC	110 to 240 VAC	100 to 240 VAC
Rated frequency	50 to 60 Hz	50 to 60 Hz	50 to 60 Hz
Rated current	4A @ 230 VACx2	5A @ 230 VACx4	12A @ 230 VACx2

Table A-9**Climatic Environment — All Models**

	<i>Operating</i>	<i>Shipping and Storage</i>
Temperature	50° to 95° F 10° to 30° C	-4° to 140° F -20° to 60° C
Relative humidity	20–80% non-condensing	15–95% non-condensing
Altitude	0 to 10,000ft 0 to 3,048m	0 to 39,370ft 0 to 12,000m

Table A-10 **Physical Specifications**

	<i>DXi3500</i>	<i>DXi5500</i>
Width (in/cm)	19 / 48.3	19 / 48.3
Height (in/cm)	3.5 / 8.8 2U	8.75 / 22.2 5U
Length (in/cm)	27 / 68.5	25.4 / 64.5
Weights (lbs/kg)	50 / 22.68	122 / 55.34
		<i>DXi7500</i>
12TB increment height (in/cm - U)		5.3 / 13.3 – 3U
24TB unit height (in/cm - U)		31.5 / 80 – 18 U
240TB unit height (in/cm - U) (two racks)		63 / 60 – 36U (x2)



Use replication to automate disaster recovery across sites!

Make a meaningful impact on your data protection and retention

What are the true costs in storage space, cooling requirements, and power use for all your redundant data? Redundant data increases disk needs and makes backup and replication more costly and more time-consuming. By using data de-duplication techniques and technologies from Quantum, you can dramatically reduce disk requirements and media management overhead while increasing your DR options.

Discover how to:

Eliminate duplicate data

Reduce disk requirements

Lower network bandwidth requirements

THE
DUMMIES
WAY

Explanations in plain English

"Get in, get out" information

Icons and other navigational aids

A dash of humor and fun

Get smart!

@ www.dummies.com

- ✓ Find listings of all our books
- ✓ Choose from among many different subject categories
- ✓ Sign up for eTips at etips.dummies.com

ISBN: 978-0-470-26054-8

Book not resalable

For Dummies®
A Branded Imprint of

