



Confidence in the connected world.

# **Global Data Deduplication Results with NetBackup PureDisk**

Storage and Bandwidth  
Optimization Data from 5  
NetBackup PureDisk Customers

*Peter Elliman, Sr. Product Marketing, NetBackup  
Platform*

*Stefaan Vervaet, Sr. Technical Product Manager,  
NetBackup PureDisk*

# Global Data Deduplication with NetBackup PureDisk

## Deduplication results from 5 customer environments

### Content

<b>Introduction</b> .....	3
<b>Where to Deploy PureDisk Data Deduplication?</b> .....	4
<b>Veritas NetBackup PureDisk from Symantec</b> .....	4
<b>Network Bandwidth Savings with PureDisk Client-Side Deduplication</b> .....	6
Understanding NetBackup PureDisk Client-Side Deduplication .....	7
Deduplication versus Compression.....	7
PureDisk Deduplication by File Type.....	8
Microsoft Exchange Server Backups Examined .....	8
Deduplication Results in the First 10 Days.....	10
<b>Bandwidth Optimization versus Storage Optimization</b> .....	11
<b>Storage Optimization from PureDisk Deduplication</b> .....	12
<b>Summary</b> .....	14
Where to get more information.....	14

## Introduction

Enterprises are seeking new ways to tackle their data protection challenges. While data growth is not new, the pace of growth has become more rapid, the location of data more dispersed, and the linkage between data sets more complex. Data deduplication offers companies the opportunity to dramatically reduce the amount of storage required for backups and to more efficiently centralize backup data to multiple sites for assured disaster recovery. Veritas NetBackup PureDisk from Symantec uses data deduplication technology to help customers address these data protection challenges.

At the heart of NetBackup PureDisk customers will find flexible deduplication technology, a highly scalable software-based storage system, and integration with NetBackup for greater functionality. PureDisk deduplication technology can be deployed into backup environments in two different places. It can be deployed at the start of the backup process by placing a PureDisk client on the server; or it can be employed, without a client, at the end of the backup process when a NetBackup media server writes data to disk. Regardless of where it is deployed NetBackup PureDisk can improve the backup process and help to accomplish the following:

- Reduce bandwidth consumed by traditional network based backups by up to 500x
- Reduce backup storage consumption by 10 to 50 times as compared to traditional tape-based backup methods

Although data deduplication technology has existed for more than five years, most organizations have yet to take advantage of the dramatic operational and storage efficiencies to be gained through deduplication. This paper examines five customer environments and presents the results achieved with NetBackup PureDisk across a mix of environments, operating systems, and applications. A sample of their environments revealed the following:

- PureDisk can reduce backup storage by 10x or greater when compared to traditional backup methods with tape media
- PureDisk deduplication reduces bandwidth by 97% or more for file servers with office data
- PureDisk can reduce storage for Microsoft Exchange PST backups by 98% or higher

## Where to Deploy PureDisk Data Deduplication?

As customers evaluate where to deploy NetBackup PureDisk data deduplication technology they seek to better understand both the bandwidth and storage efficiencies that can be derived when using this technology.

- Where can I use PureDisk data deduplication technology?
- What type of bandwidth savings can occur with PureDisk client-side deduplication?
- What is the typical rate of deduplication for different file types, operating systems, and application such as Microsoft Exchange?

## Veritas NetBackup PureDisk from Symantec

Veritas NetBackup PureDisk from Symantec offers customers bandwidth and storage optimized data deduplication as well as integration ties with NetBackup and Backup Reporter. As part of the NetBackup platform, PureDisk enables customers to centralize and manage distributed data, optimize storage of backup data on disk, and reduce their rotation and use of tape for disaster recovery.

Data deduplication technology can be deployed to optimize bandwidth and storage usage, using client-based deduplication, or just storage using target-side deduplication.

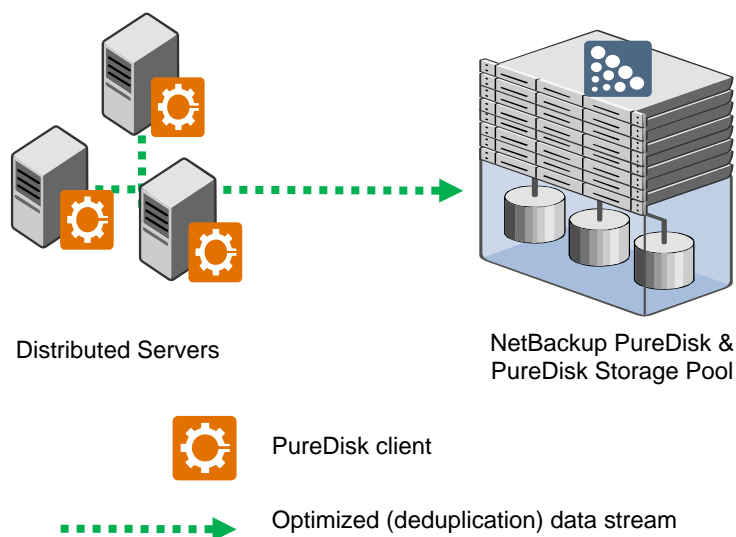


Figure 1 - Client-side deduplication

In order to accomplish client-side deduplication, a client (or agent) must exist on the system to be protected. This client enables data to be reduced before it ever begins its journey towards a backup target. Most customers deploy these agents to systems in environments with limited bandwidth or throughput capacity, typically remote office, distributed servers in a mid-size office, or a virtual server (i.e., guest OS).

If bandwidth constraints are not an issue for the backup process, as is typical in data center environments, then some customers prefer to deploy a target-based deduplication system because this typically requires the least amount of change to an existing backup architecture. There are two types of target-based deduplication systems – hardware based (appliances) and software based. With NetBackup PureDisk 6.5 (available in early 2008), Symantec will offer a software-based deduplication storage system, designed specifically for NetBackup, that offers customers the ability to store their backup data, anywhere, on any type storage. With PureDisk deduplication, customers can not only keep more backup data on disk in the data center, but leverage the bandwidth efficiency to replicate the data to a disaster recovery site.

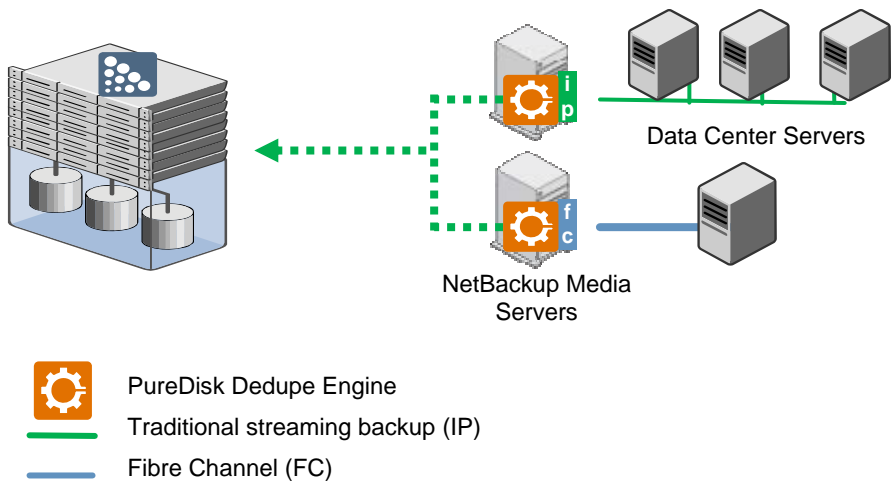


Figure 2 - Target-side deduplication (available in 2008)

### Network Bandwidth Savings with PureDisk Client-Side Deduplication

NetBackup PureDisk deduplication dramatically reduces the amount of network bandwidth required for backups. The data deduplication technology can reduce the size of backups across Windows, Linux, and UNIX operating systems for 1 or 1000s of systems and can be used in environments with network speeds ranging from 128kb/s to 1MB/s. The following table illustrates both the “reduction level,” the size of a daily PureDisk backup versus a traditional full backup, and the “reduction factor” as a measure of how many daily PureDisk backups it would take to equal one full backup.

Customer	Industry	Source Data Type	Source Amount (GB)	Daily Transfer Amount (GB)	Reduction Level	Reduction Factor
A	Public Sector	Microsoft Windows 2003 file servers - documents only	714	1.50 (0.21%)	99%	476:1
B	Consumer Goods	Microsoft Windows 2003 & UNIX file servers with office documents	2,683	20.72 (0.77%)	99%	130:1
C	Financial Services	Office files, Microsoft Exchange, and Microsoft Windows Registry (20%)	1,073	22.76 (0.21%)	97%	47:1
D	Public Sector	Microsoft Windows 2003 File Server	4,332	11.23 (0.26%)	99%	386:1
E	Oil & Gas	Microsoft Windows 2003 File Servers - documents only	848	1.14 (0.13%)	99%	744:1

Number of clients / servers between 10 and 300. Retention times range from: 11 to 331 days

**Table 1 – NetBackup PureDisk Bandwidth Optimization for 5 customers**

## Understanding NetBackup PureDisk Client-Side Deduplication

Most customers perform a full backup once per week and an incremental backup every other day of the week. With PureDisk client-side deduplication this concept disappears. NetBackup PureDisk requires only one initial full-backup after which it identifies new or changed files using a data fingerprinting algorithm, rather than a file attribute such as the modification date (as used in a traditional backup process). This approach allows for very small daily backups with the ability to recover a full image from any backup on any day.

In Table 1, **Source Amount** represents the total amount of data across all systems that must be protected each day. In a typical backup cycle all of this data would be transmitted across the network each week when performing a full backup. NetBackup PureDisk client-side deduplication eliminates this network burden. The **Daily Transfer Amount** represents the average amount of backup data moved for each day. Since the concept of duplicate full backups disappears with client-side deduplication, this represents the amount of data moved for both a “full” and an “incremental” backup.

We use two different metrics to communicate the results of deduplication because they impact users differently. As noted earlier, the **Reduction Level** metric provides an easy and relative metric for the reduction in daily bandwidth and storage required for a backup (1-(1.5 GB / 714 GB) = 99%). The Reduction Factor provides this information as a factor of source volume (714 GB / 1.5 GB = 476:1) and quickly conveys how many daily PureDisk backups would be needed to equal the bandwidth consumption of what was once a weekly full backup. Finally, these customers had a broad range of servers protected and retention times that varied anywhere from 11 to 331 days.

## Deduplication versus Compression

Customers often want to compare compression and deduplication technologies because both can reduce the size of a backup. However, the process and results of each approach differ dramatically. Compression can be applied to files, directories, or even volumes, but it lacks awareness of the underlying data and therefore can not recognize that two identical files exist in different directories. More importantly, a compression algorithm, unlike the deduplication process, can neither recognize changed data nor capture these unique blocks at a sub-file level. In other words compression reduces the size of data that it encounters, but lacks global content awareness, which impacts the space used to store this data.

For example, if a customer backed up data on a remote server and chose to compress that data before backing it up, they might achieve anywhere from 2:1 to 3:1 compression. Thus a

smaller file would be sent to the local backup application or across the network to a centralized backup application. With NetBackup PureDisk deduplication, the file on this remote server might be recognized as an identical file to one already stored, before compression occurs, in which case only a very small file (the fingerprint) would be transmitted and stored. Any system sending data to the same PureDisk Storage Pool will have awareness of any similar information that may have already been transmitted and stored. Finally, in addition to deduplication, PureDisk also offers the option to compress data the unique data segments that it identifies with a file.

### PureDisk Deduplication by File Type

The change rate of data and file types can affect the level of deduplication. The deduplication process is most effective with uncompressed file formats. Just as the mileage of a car varies based on the weather conditions and the road quality, the effectiveness of deduplication will vary based on the type of source data and the change rate of that data. The following table provides a guideline for PureDisk deduplication results based on common data types.

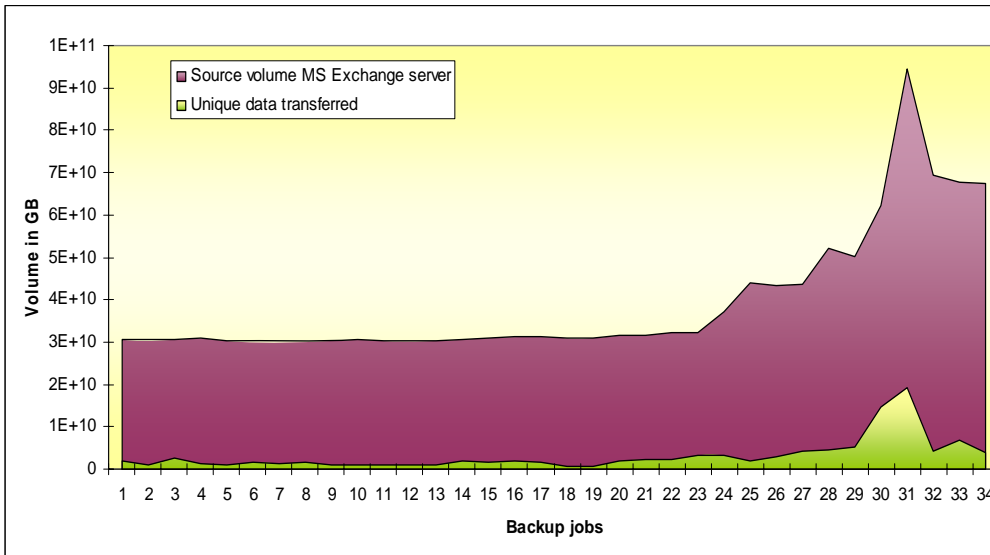
Type source data	Deduplication optimization range	Daily transferred data (PureDisk Client)
Microsoft Exchange (PST)	98% - 99.8%	0.25% - 2.0%
Office documents	98% - 99.8%	0.25% - 2.0%
Microsoft Exchange Server 2003	90% - 99.5%	0.5% - 10%
System State	80% - 90%	10% - 20%

Table 2 - Deduplication efficiency by data type

The **Deduplication Optimization Range** represents the amount of source volume data not transferred following deduplication (the higher, this value, the less data transmitted and stored each day). Thus in *Table 2 - Deduplication efficiency by data type*, backups of office documents typically reveals 98 to 99% duplicate data after the initial backup.

### Microsoft Exchange Server Backups Examined

Changes to application data can occur in surprising and unpredictable ways. The following example from one customer with Microsoft Exchange server illustrates the impact that server growth can have on an average backup environment and demonstrates how PureDisk can reduce the impact of these changes. This is important because applications like Microsoft Exchange control unstructured data that frequently changes size. These fluctuations can temporarily impact optimization and are sometimes unpredictable.



**Figure 3 - Example MS Exchange server source volume growth and corresponding transferred data**

Figure 3 - Example MS Exchange server source volume growth and corresponding transferred data depicts transferred Exchange server data (green) versus original source volume data (purple) for 34 days. As noted in Table 2, the daily transferred data amount (as a percentage of protected source data) ranges widely between 0.5% and 10%. A closer examination shows how unexpected environment changes can also affect these metrics.

A stable period persisted for an initial 23 days. During that interval, the protected source data amount and daily transferred data amounts stayed relatively constant with the average amount of daily transferred data constantly hovering around 0.5%. From day 24 to day 30 the customer added more Exchange server mailboxes which resulted in a surge in source data to protect and a corresponding increase daily transferred data. The surge was temporary however (as shown in the graph peak). After the 33rd day, the source volume and daily transferred data declined to new steady-state averages. PureDisk could accommodate these challenges and minimize the immediate impact on IT systems.

### Deduplication Results in the First 10 Days

PureDisk deduplication technology improves with a wider pool of clients. When using client-based deduplication, the benefit that one client receives from others in its pool can be seen immediately. As more PureDisk clients access the same data pool, each successive client benefits from the data sent by previous clients. These benefits can begin occurring as early as the first backup and grow overtime. We compared the initial backup for a series of clients to subsequent backups over a period of 10 days in *Figure 4 – Global deduplication benefits with 10 NetBackup PureDisk clients*.

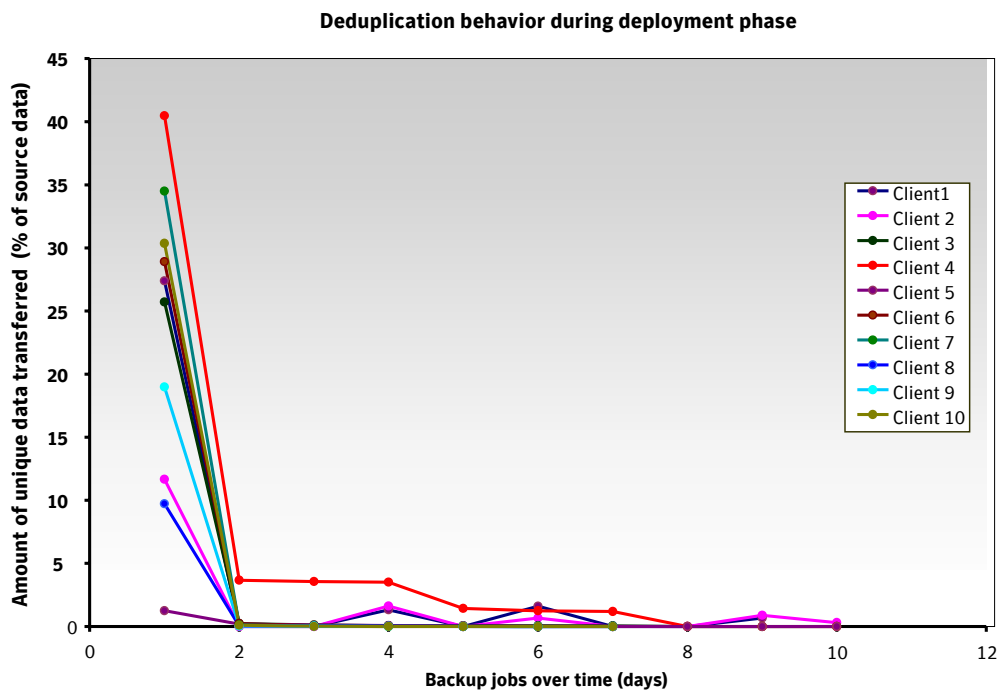


Figure 4 – Global deduplication benefits with 10 NetBackup PureDisk clients

Duplicate data was eliminated beginning with the first full backup. The y-axis illustrates that the amount transferred data during the initial backup had a range of 10% to 40% the amount of the original source data. In other words, between 60% and 90% of the data across these 10 clients was duplicate data that did not need to be sent across the network or written to backup media. By day five or six, every backup client achieved a very high optimization level (relative to a full uncompressed backup).

## Bandwidth Optimization versus Storage Optimization

A frequent point of confusion in any deduplication discussion revolves around the difference between bandwidth and storage savings. Client-side deduplication approaches deliver bandwidth savings and storage savings, while target-side systems focus only on storage reduction. Thus, a comparison of a traditional streaming backup to one done with client-side deduplication, will yield higher “deduplication savings” if one compares the bandwidth savings over a given retention period. These bandwidth savings makes NetBackup PureDisk clients ideal for the protection of distributed servers or virtual machine guests with bandwidth (and I/O) constraints.

Backup Type	Traditional Streaming (% of source data)	PureDisk Client Deduplication (% of source data)	Bandwidth Reduction Factor vs. Traditional
Initial Backup (Full)	100%	25%	4x
Incremental Backups	12%	1%	12x
Subsequent Full Backup	100%	1%	100x

Table 3 - Comparison of average volume of data by backup type

A traditional full streaming backup sends 100% of its data across the network to the backup application. The backup application then directs the data to a media source, normally tape or disk, where it can be compressed (or deduplicated). A NetBackup PureDisk client, installed on the system to be protected, deduplicates the data before it moves across the network. As the *Table 3 - Comparison of average volume of data* illustrates an initial full backup done by a PureDisk client typically requires 4x less bandwidth.

Incremental backups made with a NetBackup PureDisk client show bandwidth improvements of 12x because PureDisk deduplication technology backs up only the changed segments of a modified file, not the entire file (as happens with a traditional streaming backup). Finally, a traditional backup approaches requires subsequent full backups, which again require a large amount of bandwidth, especially versus PureDisk client-side deduplication which typically reduces in 100x less data to move across the network. Of course, some customers can reduce bandwidth consumption from weekly full backups by creating “synthetic backups” (NetBackup offers this feature), a process that assembles a new full backup based on previous incremental backups, rather than a new full backup. However, the synthetic backup approach lacks the high degree of data reduction that comes from PureDisk deduplication.

## Storage Optimization from PureDisk Deduplication

Both client and target-side deduplication approaches deliver identical storage optimization levels, but they do so at different points along the backup data path. The later approach performs deduplication before the data is written to disk. The degree of storage optimization that a customer achieves depends on the type of data, the change rate of data, and the retention time. The largest amount of optimization comes from traditional office file types. The least amount of optimization comes from compressed data formats such as those used for music, video, and medical imaging. Finally, as retention time increases, the benefits from deduplication will increase when compared to traditional backup approaches.

The following series of tables compares traditional backups with client-based deduplication found in NetBackup PureDisk for a public sector customer (customer A). The results show the following:

- Reduction in bandwidth for a full backup: 476x
- Reduction in storage versus tape (compression) over retention time: 14x

The first table (*Table 4 - Bandwidth Optimization from Customer A*) provides an overview of the customer environment and the bandwidth savings they derived with NetBackup PureDisk client-side deduplication. The important results here show the total amount of daily data transferred and stored for these systems dropped to 1.5 GB/day.

Customer	Industry	Source Data Type	Source Data (GB)	Daily Transfer (GB)	Reduction Level	Reduction Factor
A	Public Sector	Microsoft Windows 2003 file servers - documents only	714	1.50 (0.21%)	99%	476:1

**Table 4 - Bandwidth Optimization from Customer A**

The next table (*Table 5 - Full and Incremental Backup Storage Comparison - Customer A*) compares the storage optimization that this customer received from PureDisk versus their approach with tape. We have broken out these savings for both full and incremental backups so that customers can better understand the benefits.

Storage Optimization by Backup Type	Data Written to Tape (with compression)	NetBackup PureDisk
Full Backup as % source	50%	22%
Full Backup Size (GB)	357 GB (714*50%)	157 GB (714*22%)
Incr. Backup as % of source	4.0%	.21%
Incr. Backup Size (GB)	29 GB (714*4%)	1.5 GB (714*.21%)

**Table 5 - Full and Incremental Backup Storage Comparison - Customer A**

The final table brings together the storage savings picture by factoring in the retention time for each data set. Customer A had a retention time of 42 days (6 weeks) and wanted a recovery point for every day. With these recovery-point objectives (RPO) the customer had 6 full backups and 36 incrementals (42 days retention). With NetBackup PureDisk they had the equivalent of 1 full backup and 41 incrementals over the same period (42 days total), though a full backup could be recovered at any time.

Aggregate Storage Optimization by Backup Type	Data Written to Tape (with compression)	NetBackup PureDisk
Full Backups Stored	2.1 TB (357GB * 6)	.15 TB (157GB / 1024)
Incr. Backups Stored	1TB (29GB *36)	.06 TB (1.5GB*41)
Total Storage Required	3.1 TB	.21TB
Storage Optimization vs. Tape with Compression		14x (3.1 TB / .21 TB)

**Table 6 – Aggregate Storage Optimization Comparison - Customer A**

We have provided both bandwidth and storage optimization numbers next to one another so that customers can better understand how to evaluate deduplication. Obviously, the bandwidth optimization derived from NetBackup PureDisk clients allows customers to centralize the protection of systems in environments with limited network bandwidth (e.g., remote offices, distributed servers, or virtual machines). And the storage optimization from PureDisk deduplication makes the storage and management of backup data on disk across multiple locations both feasible and cost effective.

## **Summary**

NetBackup PureDisk technology enables both client-based data deduplication to optimized bandwidth usage and target-based deduplication for storage optimization. For customers with distributed servers in remote offices, Symantec offers NetBackup PureDisk, a stand-alone solution that uses client-based deduplication to deliver bandwidth efficient, storage-optimized data protection for distributed data in remote offices, data centers, and virtual environments. For systems within the data center, customers can leverage target-based deduplication, integrated with NetBackup, by using the PureDisk Deduplication Option. Regardless of where customers choose to deploy NetBackup PureDisk technology, Symantec software allows customers to build a scalable deduplication storage system using any combination of servers and storage. While data reduction results may vary based on data type, change rate of data, and the number of servers, the opportunities for management and storage cost efficiencies with PureDisk technology remain compelling.

## **Where to get more information**

Additional whitepapers and other information about NetBackup PureDisk can be found by visiting our product pages on the Symantec website at [www.symantec.com/puredisk](http://www.symantec.com/puredisk).

You may also contact Symantec Business Sales in the United States at: 800-745-6054. For phone numbers outside the United States, please visit our corporate website at [www.symantec.com](http://www.symantec.com).

## About Symantec

Symantec is a global leader in infrastructure software, enabling businesses and consumers to have confidence in a connected world. The company helps customers protect their infrastructure, information, and interactions by delivering software and services that address risks to security, availability, compliance, and performance. Headquartered in Cupertino, Calif., Symantec has operations in 40 countries. More information is available at [www.symantec.com](http://www.symantec.com).

For specific country offices and contact numbers, please visit our Web site. For product information in the U.S., call toll-free 1 (800) 745 6054.

Symantec Corporation  
World Headquarters  
20330 Stevens Creek Boulevard  
Cupertino, CA 95014 USA  
+1 (408) 517 8000  
1 (800) 721 3934  
[www.symantec.com](http://www.symantec.com)

Copyright © 2007 Symantec Corporation. All rights reserved. Symantec and the Symantec logo are trademarks or registered trademarks of Symantec Corporation or its affiliates in the U.S. and other countries. Other names may be trademarks of their respective owners.

11/07 13561423