

Symantec NetBackup PureDisk

Optimizing Backups with Deduplication for Remote Offices, Data Center and Virtual Machines

*Mayur Dewaikar
Sr. Product Manager
Information Management Group*

Symantec NetBackup PureDisk

Optimizing Backups with Deduplication for Remote Offices, Data Center and Virtual Machines

Contents

Executive Summary	1
Overview of PureDisk Deduplication Technology	1
PureDisk Use Cases	2
Remote Office Backups	3
Data Center Backups	4
Virtual Machine Backups	6
Disaster Recovery with PureDisk	7
Key PureDisk Benefits	9
Conclusion	9

Executive Summary

Most enterprises today are challenged with the problem of rapid data growth. Both IDC and the Enterprise Strategy Group (ESG) studies predict the data growth to be in the 50-60% per year range. As the enterprises continue to grow via mergers and acquisitions, the data is no longer confined to a single data center, but is often spread across multiple data centers, remote offices and even virtual machine environments.

With the heavy reliance of businesses on IT systems, IT managers are under tremendous pressure to decrease downtime and deliver more stringent recovery point objectives (RPOs) and recovery time objectives (RTOs).

While data continues to grow at a rapid pace, backup windows have either stayed the same or shrunk. Add to that shrinking IT budgets and you quickly realize that adding more storage is no longer a viable solution for addressing the data protection needs of a growing enterprise. A thorough reassessment of current data protection architecture, technologies, and processes must take place to arrive at the right solution.

As part of this data protection infrastructure reassessment, many companies across the globe have now begun a transition from tape to disk-based data protection. Tape systems have been notorious for their inadequacy especially around performance, recoverability, and reliability. Disk-based backups solve many of the challenges faced by tape backups, but they only solve part of the overall data protection challenge especially when used with traditional backup products.

Traditional backup products generally require a rotational schedule of full and incremental backups, which results in significant amount of data movement on a weekly basis. Despite the decline in cost of disk storage, companies soon realize that not all data can be stored on disk for local recovery.

Symantec's NetBackup PureDisk, with built-in global deduplication technology, makes disk backups more cost efficient by eliminating the backup of duplicate data while addressing the problems associated with traditional backup products. NetBackup PureDisk offers customers a scalable software-based data deduplication solution that integrates with NetBackup and provides customers with the critical features required to protect all their data - from remote office to virtual environment to the data center. It reduces the size of backups with a global deduplication technology that can be deployed for storage reduction, using integration with NetBackup, or for bandwidth reduction using PureDisk clients. Data backed up via PureDisk is encrypted in-flight and also at rest and can be made highly available using integration with industry standard HA solution, Veritas Cluster Server, from Symantec. Built-in replication allows independence from tape for DR purposes. Finally, an open architecture allows customers to easily deploy and scale NetBackup PureDisk using standard storage and servers.

Overview of PureDisk Deduplication Technology

PureDisk offers segment level global deduplication for the enterprise. During the backup process, the backup data set is broken down into smaller segments and each segment is assigned a hash value which is calculated based upon the binary content of a file. This is done so as to uniquely identify the data segments, rather than depending on the file path and name on any given hardware device. Since the sequence can be used to uniquely identify a file by its contents, it is called

the fingerprint. The system therefore refers to files in the same manner that the Internet refers to servers. Moreover, files become referable regardless of their position on a given device. The fingerprint is derived from the total contents of the file. The result is that files with the same content will have the same fingerprint, even when the files have different names, locations, attributes, creation or modification dates, and security attributes. Files with different content will lead to a different fingerprint. Indeed, only a comparison of two fingerprints is required to know if two files with different metadata (filename, path name, etc.) are unique or not.

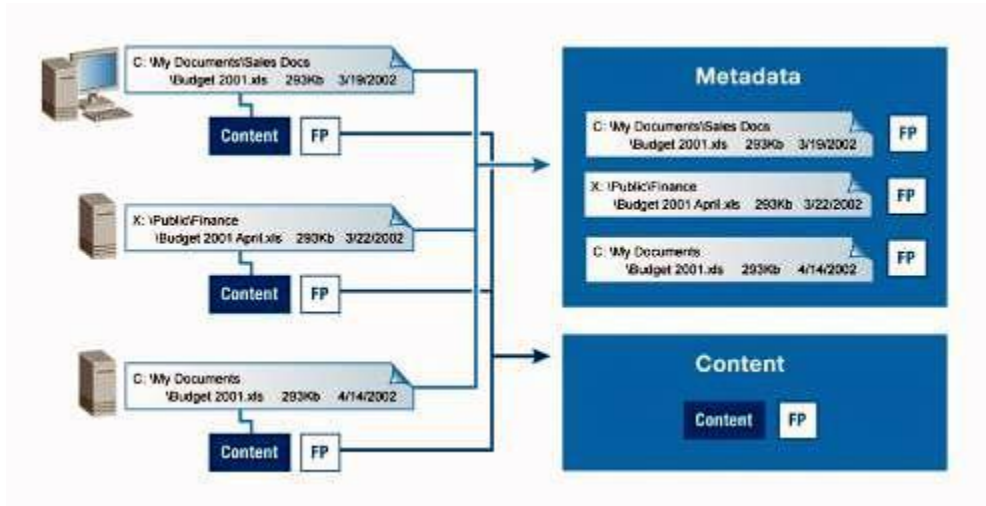


Fig 1: PureDisk Deduplication Process

PureDisk Use Cases

PureDisk can be used in three use cases:

1. Remote Office Backups
2. Data Center Backups
3. Virtual Machine Backups

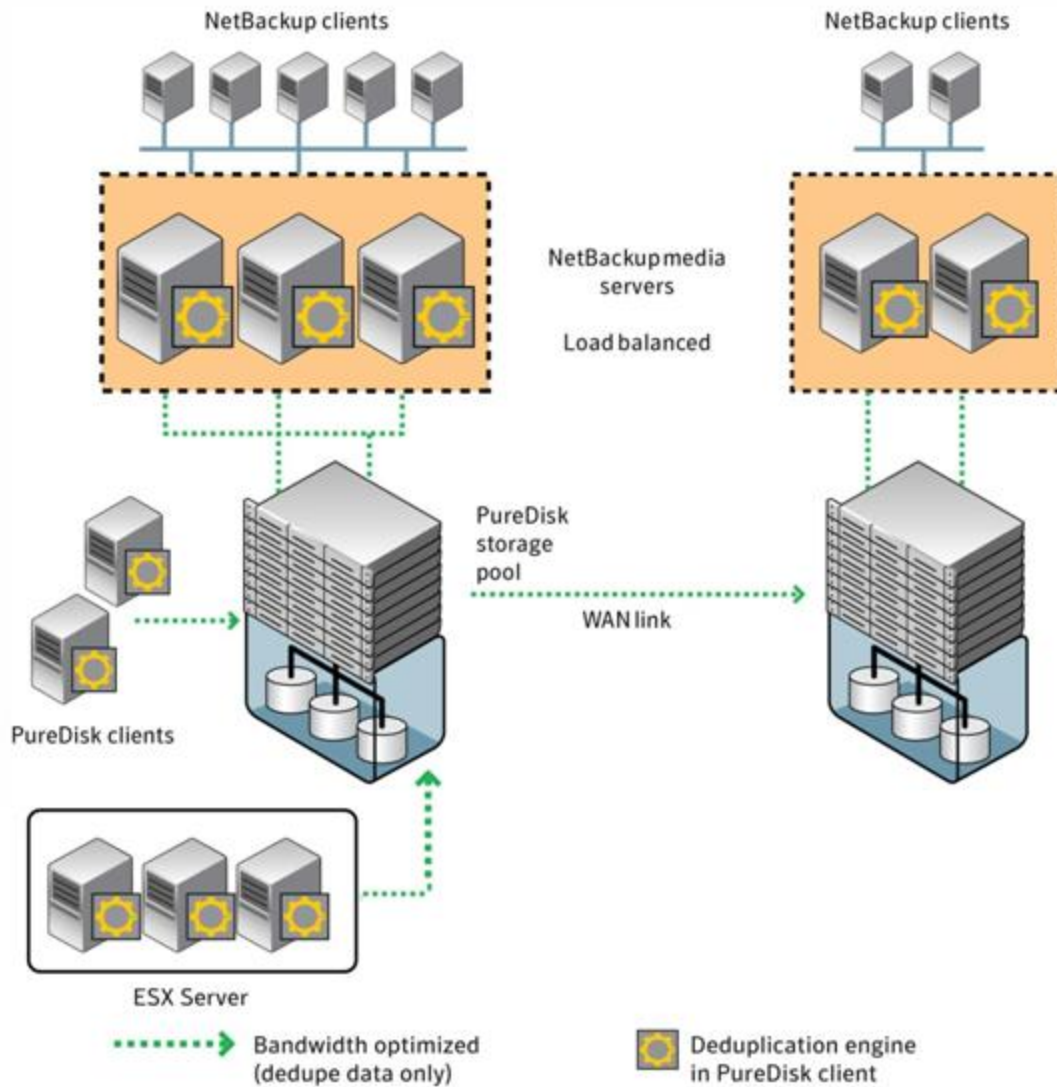


Fig 2: Typical PureDisk Architecture

Remote Office Backups

The globalization of businesses and cultural changes such as home and mobile working has made the backup process more complex. Companies operating from different local offices however, must distribute parts of their IT infrastructure over these remote sites out of necessity. Local documents, emails, presentations, etc. are kept on local file servers primarily to improve network performance and to allow rapid recovery in the event of data loss. Many enterprises organize the backups of their remote sites locally, rather than sending data over a WAN network to a central backup server. This practice, however, raises several important issues:

- Can administrators guarantee that backups are being performed according to policy or even performed at all?
- Are the backups successful?
- Are skilled resources available to troubleshoot errors?
- Are the tapes stored securely, protected from potential harm or loss?

PureDisk helps eliminate these bandwidth and tape-related issues by combining sophisticated disk-based backup with data deduplication. PureDisk works at the *source* to eliminate data redundancy *before* it traverses the network and enters the data center.

In case of smaller remote offices with fewer servers, the PureDisk agent is installed on the remote machines and the data is generally directly backed up to a PureDisk storage pool in a central data center. If the remote office is large with several servers and significant amount of data, a local PureDisk storage pool can be deployed at the remote site which can then replicate data back to the central data center. A local instance of PureDisk allows for faster backups and recovery. When compared with a traditional full backup over a typical retention period, PureDisk can reduce storage consumption by up to 500x and bandwidth consumption up to 50x.

With such significant bandwidth and storage savings, Symantec expects that organizations can:

- Cost effectively substitute tape-based backups with centralized disk-based backups.
- Shorten the time required for backup as less data needs to be backed up each day.
- Reduce or eliminate the administrative burden of managing backup applications and tape devices in remote offices.
- Eliminate tape handling and offsite shipping & storage costs.

Data Center Backups

In the data center, the biggest challenge the backup admin often runs into is with the enormous amounts of data that needs to be backed up on a daily basis within a backup window that stays constant or keeps shrinking. Aggressive RTOs also mean that the data should be recovered with minimal down time. In the past, tapes were the de facto standard for storing backup data, but in this day and age, tape backups are unlikely to meet the needs of a growing enterprise. It is therefore no surprise that a vast majority of companies are now exploring some form of disk backups in their environment. However, the flexibility, performance, and scalability of disk backups come at a cost that needs to be carefully managed.

In the data center use case, using the PureDisk deduplication option (PDDO) for NetBackup, NetBackup is able to use PureDisk Storage Pool as an intelligent deduplicated disk target. This option is different in that the deduplication is performed in software on the media server, rather than on the target device. Deploying disk arrays as a backup target for NetBackup requires a lot of capacity as every backup (full or incremental) is stored as a backup image, which potentially represents a lot of redundant data. When NetBackup is used with PDDO, it allows customers to more efficiently use their disk capacity by eliminating the redundant data and thus allowing more versions of the data to be retained on disk for longer periods of time. More data available on disk means better supportability for the established RTOs and RPOs.

Symantec NetBackup PureDisk
 Optimizing Backups with Deduplication for Remote Offices, Data Center and Virtual Machines

In the PDDO use case, PureDisk offers customers a low-cost, scalable, and high performance alternative to appliance based target deduplication devices. Storage and performance scale independent of each other. Each PureDisk environment allows customers to protect up to a PB of data on 100TB of dedupe backend capacity. The storage capacity of the environment can be scaled via the addition of PureDisk nodes and requires absolutely no down time. Target deduplication devices do not offer this flexibility and are limited in scalability as they typically only deduplicate within the device, not across devices. This requires backups from a particular client to be stored to the same device at all times to benefit from deduplication, which then limits the flexibility of using NetBackup load balancing, capacity management, backup spanning, and failover features.

PureDisk Deduplication Option deduplicates the NetBackup data streams on the NetBackup media server. Multiple deduplication processes can run on parallel backup streams on a media server, and multiple media servers can share a single PureDisk storage pool. This architecture is set up to scale the deduplication performance beyond the boundaries of a single processing head while maintaining deduplication across the whole dataset. With appliances; deduplication only happens when data enters the appliance, target deduplication devices are limited by the aggregate deduplication performance of the device.

The following table gives a flavor of the throughput characteristics of the PureDisk Deduplication Option when used with NetBackup.

PureDisk Deduplication Option Backup Throughput			
# of PureDisk Nodes	# of NBU Media Servers	Usable Capacity	Maximum Throughput
1	2	16TB	2.5TB/Hr
2	2	32TB	2.8TB/Hr
4	4	64TB	3.5TB/Hr

Assumes multi-stream NBU backups
 Backup data set consisting of typical enterprise mix (file and folder, databases, mail)
 Best possible performance achieved over 10GigE Ethernet
 Hard drives configured in a RAID 6 configuration

Fig 3: PDDO Performance Throughputs

To improve DR processes, the built-in replication capabilities of the PureDisk Storage Pool can be controlled from NetBackup to create an off-premise copy. In addition, the PureDisk Storage Pool nodes can be configured as a cluster for High Availability. With PureDisk storage as a target for NetBackup, IT has more choices. In addition to allowing flexibility for disk type, vendor, and capacity, IT has the ability to set inline or post-process deduplication strategies on a per backup policy basis (client-side inline for distributed backup, inline for LAN rate backup, and post-process for SAN rate backup, for example). Data deduplication on the media server allows throughput to be scaled with the number of media servers. Importantly, PureDisk Storage Pool, as its name implies, allows for storage to be pooled and deduplication to occur across remote office, virtual machine, and data center backup sets.

Virtual Machine Backups

IT organizations have discovered that virtualization technology can simplify server management and reduce total operating costs. Although virtualization brings a lot of benefits, it introduces some new challenges too. A typical challenge:

- When you're running your servers in a virtual environment, it's really easy to create new virtual machines. It's so easy, in fact, that it creates what they call "VM Sprawl", which increases the management cost. With the growth in the number of virtual machines comes the growth in storage needed to host the virtual machines.

From a data protection standpoint, protecting virtual machines can be more challenging than protecting physical machines. Not only does the backup application need more storage for protecting the virtual machines, but it is also competing with the virtual machines for the same shared resources on the host.

The challenge created by the rapidly growing VMware environments can be overcome by integrating technology within the backup process that will perform capacity optimization, more commonly called data deduplication.

PureDisk quickly and effectively protects virtual machines by reducing the size of the backup data across virtual machines. PureDisk also eliminates the traditional backup bottleneck caused by large amounts of data that must pass through same set of shared resources on the host such as the Ethernet adapter, CPU, memory, and disk resources.

PureDisk can be deployed to enhance VMware backups in three functional methods:

1. **PureDisk Client in each guest OS:** In this approach, the PureDisk client is directly deployed on the virtual machine. This is no different than installing a backup client on a physical host. Using the source side deduplication functionality available in the PureDisk client, the duplicate data is identified at the source itself and never backed up. As a result of this, the actual data being backed up is completely unique which reduces the impact on the ESX infrastructure resources compared to traditional backup approaches. Also, the backups are up to 10 times faster as only the unique data is transferred.
2. **PureDisk as a Disk Storage Unit for NetBackup integrating with VCB:** In this approach, the PureDisk storage pool is used as a destination for NetBackup to store full VMDK backups and file level backups captured through VMware's vStorage and VCB API. Once the virtual machines are streamed to the NetBackup Media server, VMDK images are deduplicated inline and stored on the PureDisk storage pool. The deduplication occurs on the NetBackup media server upon which only the unique data is sent over to the PureDisk storage pool for retention. This method is very effective in terms of eliminating the load of the data backup process on the ESX infrastructure and reducing the storage footprint of backups via data deduplication. Additionally, it offers two options in terms of recovery- full image level recovery and a more granular file/folder level recovery from within the single virtual machine backup. This single pass backup approach, with two recovery options, offers flexibility to customers and also significant cost savings in terms of storage.

3. PureDisk Client on the proxy server utilizing VCB: In this approach, the PureDisk client is deployed on the VMware proxy server. This enables backup of VMware guests with little impact on the individual guests or the ESX server. In this configuration, no backup software resides in the guest OS. Once the virtual machines are mounted on the VCB proxy server, they can be easily backed up by the PureDisk client that is installed on the VCB proxy server. While this approach is viable, it is not broadly considered for deployment given that option (2) stated above offers superior functionality and tends to be the preferred method of implementation.

Disaster Recovery with PureDisk

PureDisk offers two convenient disaster recovery options:

1. Built-in disk to disk replication
2. DR backup to NetBackup with tape out capability

Disk to disk replication is built into the product and is available at no additional cost. The data is replicated in a deduplicated format and is encrypted during transit and at rest. The replication option requires a PureDisk setup at the secondary. If PureDisk is being used as a target for NetBackup backups in the data center use case, the replication is controlled by NetBackup via the optimized duplication process. NetBackup remains fully aware of the secondary copy which allows for easy recovery of data in case of disasters. Disk to disk replication allows complete independence from tape, but if a customer is looking to keep a copy of the data on tape for DR purposes, PureDisk offers DR to tape via its integration with NetBackup. With this option, the data that has been backed up to a PureDisk storage pool can be sent to tapes via NetBackup. The process is fully automated via policies defined in the product GUI.

High Availability with PureDisk

PureDisk offers industry standard, low cost, built-in N+1 HA with its integration with Veritas Cluster Server. Only one standby node is required to protect all active PureDisk nodes. The Veritas Cluster Server license is included at no additional charge for use with PureDisk. When PureDisk is used with NetBackup in the data center use case, PureDisk can also leverage the NetBackup Media Server load balancing option which ensures high availability of the NetBackup environment.

Enterprise Class Backup Reporting

Veritas Backup Reporter, from Symantec provides a comprehensive reporting platform for PureDisk. Through an architecture where data is collected remotely and non-intrusively, a report portfolio ranging from operational reports enabling backup administrators to have complete visibility of what is going on across their PureDisk environment to business level reports and management roll-ups is available. The PureDisk reporting is seamlessly integrated with all other data protection reporting providing for a common and consistent user experience. There are 70+ standard pre-built reports that are a click away. These include reports on trends, forecasts, rankings (largest client, slowest client, throughput etc.) and status (success/failure) along with a core chargeback capability. Also included are reports on

Symantec NetBackup PureDisk Optimizing Backups with Deduplication for Remote Offices, Data Center and Virtual Machines

deduplication providing in depth statistical analyses around the effectiveness of the deduplication process. With a highly customizable platform, Backup Reporter provides the relevant content and context to administrator, architect, service provider, and end-user.

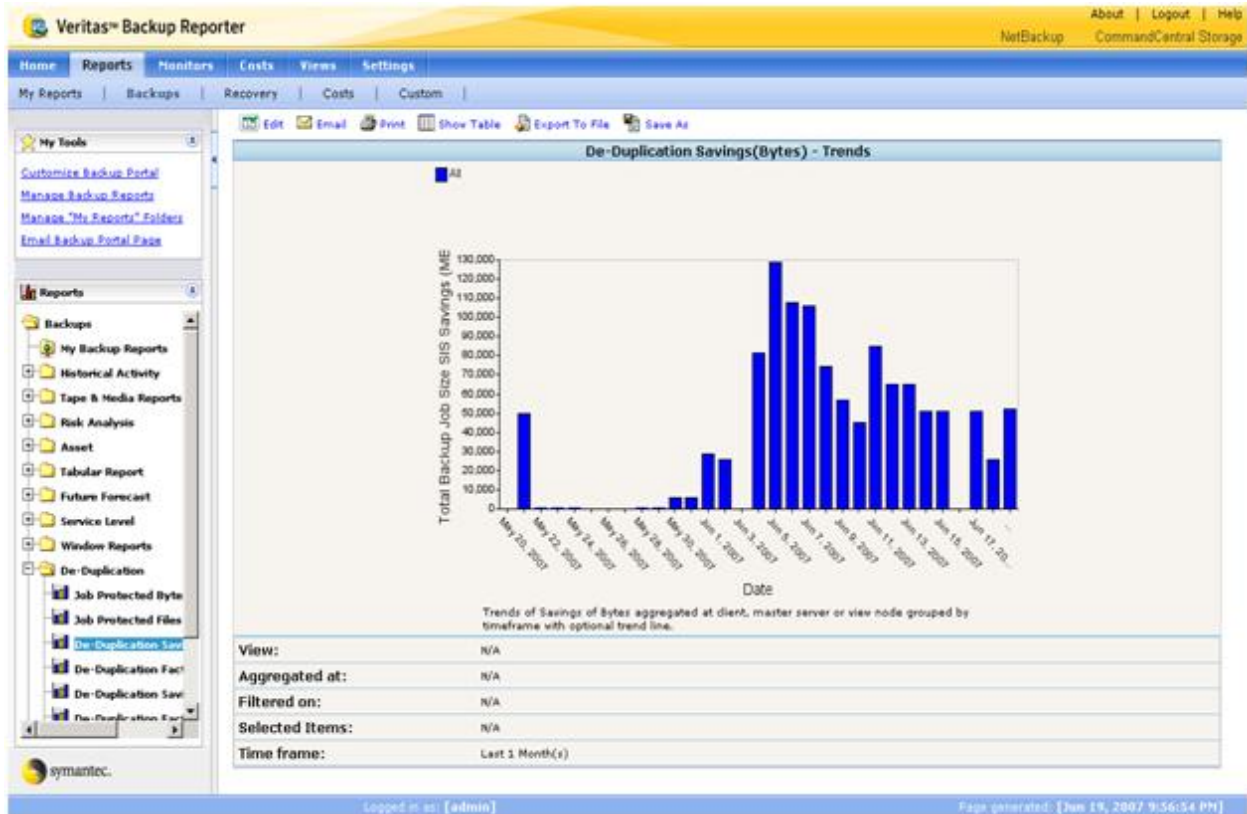


Fig 4: Sample Backup Reporter Report- Deduplication Savings

Key PureDisk Benefits

1. Integration with NetBackup media server for scalable, high performance, load-balanced disk-backup
2. Global deduplication across remote offices, data center, and virtual machines all in one single deployment
3. Up to 500x reduction in storage and up to 50x reduction in bandwidth
4. Up to a petabyte of data protected in one single storage pool
5. High scalable with up to 16 TB of usable dedupe capacity per node and 100 TB of usable dedupe capacity per PureDisk setup
6. Backup throughput as high as 3.5 TB/hr
7. Hardware agnostic approach allows any commodity server or storage to be used with PureDisk which helps to keep the total cost of ownership low
8. Built in high-availability via integration with Veritas Cluster Server at no additional charge
9. Built-in replication that allows for a tapeless DR strategy
10. Choice of in-line or post-process deduplication
11. Choice of source or target deduplication
12. Built-in 256 bit blow fish algorithm that ensures security of data in flight and at rest on the PureDisk server

Conclusion

Symantec's NetBackup PureDisk, with built-in global deduplication technology, makes disk backups more cost efficient by eliminating the backup of duplicate data thus also addressing the problems associated with traditional backup products. NetBackup PureDisk offers customers a scalable software-based data deduplication solution that integrates with NetBackup and provides customers with the critical features required to protect all their data - from remote office to virtual environment to data center. It reduces the size of backups with a global deduplication technology that can be deployed for storage reduction, using integration with NetBackup, or for bandwidth reduction using PureDisk clients. Data backed up via PureDisk is encrypted in-flight and also at rest and can be made highly available using integration with industry standard HA solution, Veritas Cluster Server. Built-in replication allows independence from tape for DR purposes. Finally, an open architecture allows customers to easily deploy and scale NetBackup PureDisk using standard storage and servers.

About Symantec

Symantec is a global leader in providing security, storage and systems management solutions to help consumers and organizations secure and manage their information-driven world. Our software and services protect against more risks at more points, more completely and efficiently, enabling confidence wherever information is used or stored.

For specific country offices and contact numbers, please visit our website.

Symantec World
Headquarters
20330 Stevens Creek Blvd.
Cupertino, CA 95014 USA
+1 (408) 517 8000
1 (800) 721 3934
www.symantec.com

Copyright © 2009 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.
9/2009 20252020