NETAPP WHITE PAPER

The NetApp Storage Efficiency Guide

Larry Freeman, NetApp
December 2008 | WP-7022-1208

USING STORAGE EFFICIENCY TO ACHIEVE END-TO-END DATA REDUCTION

Table of Contents

1 EXECUTIVE SUMMARY

As we conclude a tumultuous 2008, there are several dynamics at work in the data center. Uncertain economic times are causing data center managers to reevaluate their budgets and reduce overall IT spending. At the same time, environmental responsibility continues to be top-of-mind for facilities managers, who are concerned with the cost and availability of power, cooling, and space required by today's IT infrastructure.

While IT managers wrestle with these concerns, data continues to grow. By some estimates, the creation of a typical business file initiates a chain of events that causes that file to be copied well over 1,000 times in its lifetime. If the file is an image of a popular entertainer or a video clip of a sports figure making a heroic play, tens of thousands of copies will be quickly distributed around the globe. How does this affect data storage in enterprise data centers? Have your users downloaded images and videos and completely forgotten about them? Do your users refuse to delete old files because "you never know if you might need them"? Are you—the system administrator—reluctant to purge data volumes because no one is quite sure who the owner of the data is and what it's used for? If you answered any of these questions in the affirmative, you are not alone. Most system administrators are grappling with the constant creep of data throughout the data center. Unfortunately, there is no convenient trash can that you can throw your data rubbish into.

There are, however, many ways to attack the problem of data proliferation. First, you could demand that your accountants, engineers, managers, technicians, and executive staff immediately delete all their old unused data files. Hmmm—that went over well, didn't it? Well, you could implement a search and classify mechanism in your data center to automatically move "stale" files to a disk archival system. This frees up room on your primary storage arrays, but all that data still resides somewhere, and it's still consuming large quantities of disk drive space.

Finally, you could take advantage of existing storage efficiency techniques to manage the growth of your data, allowing you to retire legacy storage systems, postpone the purchase of a new storage system, or purchase smaller storage systems. This paper is a guide to help you understand how NetApp can enable maximum storage efficiency that will allow you to store all your data and accommodate rapid data growth without straining your people or your budget.

2 STORAGE EFFICIENCY OVERVIEW

Simply stated, storage efficiency enables you to store the maximum amount of data in the smallest possible space (and at the lowest overall cost). NetApp® Snapshot™ technology, introduced in 1992, was the first widely used storage efficiency feature available on disk-based open systems enterprise storage arrays. Snapshot copies enabled system administrators to create many point-in-time copies of their entire data volumes but consumed only a fraction of the space that would normally have been required to make backup copies of these volumes.

Snapshot copies were a disruptive technology; they caused a change in behavior of system administrators by allowing them to back up their volumes more frequently than ever before. Once per minute, once per hour, once per day—it didn't matter, because these backups were simply virtual copies that consumed very little disk space.

Today, over 15 years later, NetApp Snapshot technology has matured into an extensive suite of virtualized tools that enable system administrators to effectively provision their primary storage, manage test and development clone copies, reduce the size of point-in-time backup copies, replicate these copies across LANs and WANs, and reduce overall volume requirements by eliminating redundant data blocks.

3 STORAGE EFFICIENCY BEGINS WITH THE CREATION OF DATA

It is nearly impossible to predict how long any data file will be retained on disk. All data begins its life on primary storage. Whether it's a database entry, user file, software source code file, or e-mail attachment, this data consumes physical space on a disk drive somewhere within your primary storage environment. The creation of data on primary storage begins a chain of inefficiencies.

One of the first problems that a storage system administrator faces is quota allocation. How much physical storage space should be assigned for each particular user or application? Knowing that an overflowing data volume has many unpleasant side effects, system administrators commonly overprovision their quotas and applications. For instance, if they believe that an application will require a single terabyte, they might decide to allocate 2TB to accommodate growth over time, or to adjust for a miscalculation of the storage space actually consumed by the application.

But what if the application does not grow as expected, or the miscalculation was on the short side? The result is wasted space—space that can't be used by any other application. By some estimates, an average 60% of primary disk storage remains unused simply because of this type of overprovisioning.

This cycle of guesswork and resulting inefficiencies is not limited to primary storage. Inefficiencies begin to propogate across all storage tiers: replication copies, backup copies, and archival copies can all suffer the same fate as primary storage: improper utilization. The following sections illustrate how NetApp addresses this problem from end to end.

NetApp set out to solve the problem of overprovisioning with a technique know as *thin provisioning.* NetApp FlexVol® technology enables users to create flexible volumes that appear to be a certain size but are in reality much smaller physically.

FlexVol technology offers substantial improvements in storage provisioning, utilization, and volume sizing. Data volumes can be sized and resized quickly and dynamically as application requirements change.

The bottom-line impact of FlexVol is a dramatic reduction in physically allocated storage. Benefits include budget savings as well as related savings in data center space, power, heat, and cooling requirements.

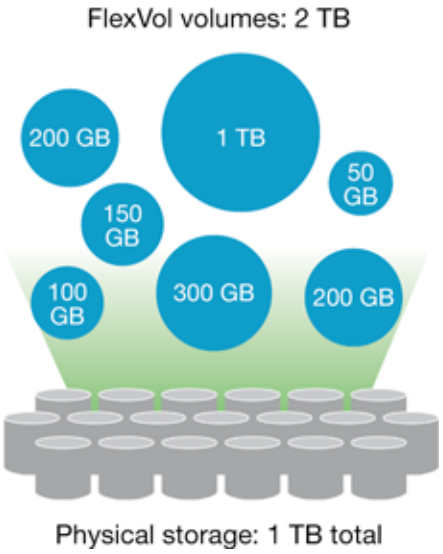

Figure 1) FlexVol allows virtual volume provisioning.

4 TEST AND DEVELOPMENT COPIES

Often, system administrators are required to allocate substantial primary storage space for essential enterprise test operations, such as bug-fix testing, platform and upgrade checks, multiple simulations against large data sets, and remote office software staging.

In addition, organizations that rely on large-scale simulations for comprehensive testing, analysis, and modeling usually incur large costs associated with provisioning additional primary storage space.

To address this issue, NetApp turned to Snapshot technology, in the form of our FlexClone® feature. Using a technique sometimes referred to as “writable” Snapshot copies, FlexClone achieves storage efficiency for applications in which temporary, writable copies of data volumes are needed.

FlexClone technology enables multiple, instant data set clones with minimal storage overhead. This is accomplished by creating a virtual copy of the primary data set and storing only the data changes between a parent volume and a clone. All unchanged data remains on primary storage and is used by both the primary application and the secondary clone copy. Multiple clone copies can be created from a single primary data set, enabling users to perform multiple simulations and to compare the characteristics of each data set after the simulations are complete.

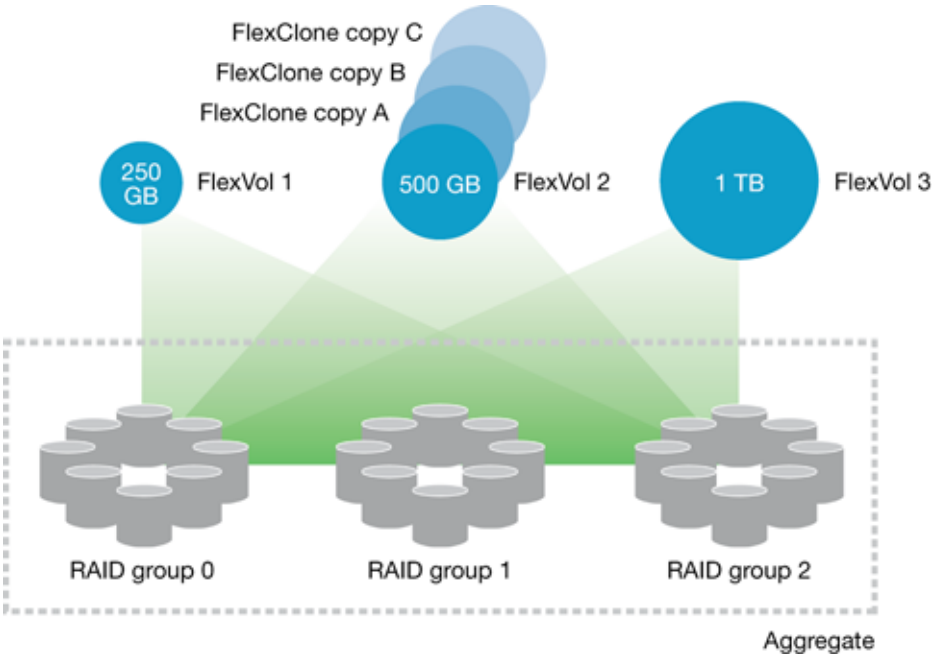

Figure 2) Using FlexClone to create virtual test and development copies.

5 EFFICIENT DATA REPLICATION

Global enterprises need to protect and quickly recover data in the event of natural and man-made disasters, operator errors, and technology and application failures. They also need a space-efficient method of distributing data to and from remote locations. Without an effective data protection and distribution strategy, operations can be brought to a standstill, resulting in millions of dollars of lost revenue.

Exceptionally powerful, yet easy to use and administer, NetApp SnapMirror® delivers the disaster recovery and data distribution solution that today's global enterprise requires. By replicating data at high speeds over a LAN or a WAN, SnapMirror software provides the highest possible data availability and fastest recovery for mission-critical applications. SnapMirror software mirrors data to one or more NetApp storage systems and continually updates the mirrored data to keep it current and available for disaster recovery.

At the core of SnapMirror software is its storage-efficient design. First, a baseline mirror is performed between the SnapMirror source and SnapMirror destination systems. Next, at user-defined intervals, Snapshot copies are taken of the source system and only the new and changed blocks are sent incrementally over the network to the destination system. When the data is received at the destination, the changed blocks are merged with the existing data blocks, resulting in a full mirror copy of the source system.

By replicating only the data that has changed since the last Snapshot copy, SnapMirror significantly reduces network bandwidth requirements. Data deduplication, described later in this paper, facilitates further space reduction at the source system, at the destination system, and during the data transfer between the two systems. The result is lower infrastructure cost for data replication and disaster recovery.

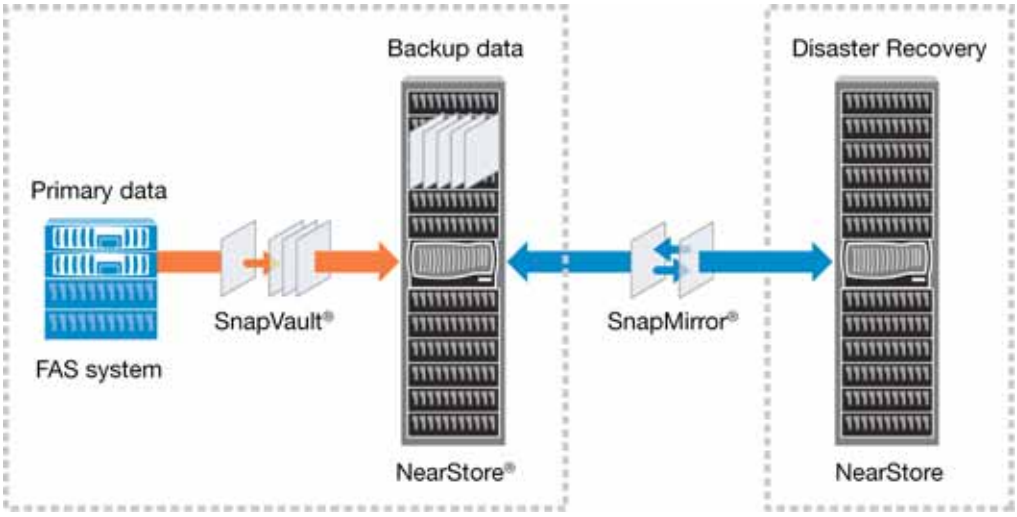

Figure 3) SnapVault and SnapMirror for space-efficient backup and disaster recovery.

USER CASE STUDY 1: FLEXCLONE

Twice a year, a leading provider of learning tools and course management systems releases upgrades of its online learning applications. To enable customers to test their customizations of these applications, the company has a shared test and development infrastructure based on NetApp storage systems. Test and development engineers use these environments to update application code, update the NetApp storage environment, test configurations, and issue security patches and fixes.

Developers use NetApp FlexClone to make working copies of the preproduction data. These cloned copies take up almost no additional storage space. NetApp Snapshot technology enables developers to make point-in-time copies of the data. "The ease of setting up a test and development system, especially with the FlexClone technology, is almost immeasurable compared to what we were doing previously," explains this customer. "It used to take 36 hours to make a working copy of our largest client's database. With FlexClone, it takes less than an hour, a reduction of more than 97%."

6 SENSIBLE DISK-TO-DISK DATA BACKUPS

Data backups are a fact of life in enterprise data centers. Like that trip to the dentist, the experience can sometimes be agonizing, but the alternatives are far worse. The longer you postpone your trip to the dentist (or attending to your data backups), the more likely it is that a much more painful event will eventually occur. NetApp helps take the pain out of data backups with SnapVault® software.

Storage efficiency again comes into play with SnapVault. First, a complete copy of the primary data is stored on the SnapVault backup system. This initial, or baseline, transfer is much like a level-zero backup to tape. Each subsequent backup, however, transfers only the data blocks that have changed since the previous backup. NetApp uses a variation of Snapshot technology to create these transfers—only the “new” data blocks are sent to the SnapVault system. This means that each subsequent backup copy consumes only an amount of disk space proportional to the differences between it and the previous backup copy. To the user, however, each backup session is virtualized to appear as though it were a complete, level-zero backup copy, greatly simplifying the process of data restoration.

For example, if SnapVault backed up a 100GB primary data set for the first time, it would consume 100GB of disk space on the SnapVault system. Over the course of several hours, suppose that users change 10GB of data on the primary file system. When the next SnapVault backup occurs, SnapVault writes the 10GB of changes to the SnapVault system and creates a new Snapshot copy. At this point, the SnapVault system contains two Snapshot copies; one contains an image of the file system as it appeared when the baseline backup occurred and the other contains an image of the file system as it appeared when the incremental backup occurred. The copies consume a combined total of 110GB of space on the SnapVault system, but they are the equivalent of two complete 100GB backup copies. System administrators can refer to any of the backup instances to retrieve their files and can easily store dozens of backup images in a reduced amount of space.

Like SnapMirror, data deduplication can be combined with SnapVault for additional savings. The result is a dramatic reduction in the physical storage requirement for disk-to-disk backups. Benefits include fast and easy restoration of files from disk and the option to retain backups on disk for longer periods of time because each subsequent backup requires very little disk space.

USER CASE STUDY 2: NETAPP DEDUPLICATION AND FLEXVOL

A major multimedia company is also a long-time NetApp customer. Among other applications, this company has three SQL servers with a total capacity of 2TB. These databases are considered essential to operations because they contain customer billing information. From the main data center, all three SQL databases are backed up nightly to a FAS270 in a second location. From that location, the three databases are again backed up to a FAS3050 in a third location for disaster recovery and archiving.

This company wanted to eliminate all tape backups and instead use NetApp for disk-to-disk backup and disaster recovery. Because of the large database size and the requirement to keep 16 backup copies online at all times, they were also interested in reducing disk space requirements.

Proof-of-concept testing with NetApp deduplication validated that 40% to 50% volume space savings would occur consistently when deduplication was performed after the second nightly backup. Once the concept was proven, an automated script was developed. All database backups were saved to FAS3050 volumes in pairs. After the second nightly database backup, deduplication is run on the volume and a check is made to determine the new (reduced) volume space required. The volume is then resized automatically using FlexVol. This process continues until 8 volumes are created, with a total of 16 database copies. After that point, on subsequent backups, the first volume is deleted and a 17th volume is created, and so on.

The results of this implementation were a completely automated database backup process and a 40% reduction in disk requirements—from 32TB to 19TB.

7 STOP BEING REDUNDANT!

Remember this well-known scene from The Sorcerer's Apprentice? A magician's assistant, having watched his master ply his craft (but not quite closely enough), casts a spell on a broom to fetch his water. Once the water is fetched, the assistant forgets the incantation to stop the broom, which soon overflows his basin and shows no sign of stopping. Desperate, the apprentice attacks the broom with an axe. Naturally (or should we say, supernaturally) the halves all come to life and continue to proliferate. Water, water everywhere! Just when all hope seems lost, the master arrives and calls the brooms to a halt.

In this story, the apprentice felt the full effect of proliferation, and was powerless to stop it. System administrators often have the same feeling of helplessness as they watch their data volume grow ever larger. As mentioned earlier, redundancy is a root cause of data proliferation. Once data is created, it seems to multiply faster than you, the apprentice, can control it. NetApp deduplication is your magic cure to stop the spread of data redundancy. The average UNIX® or Windows® enterprise disk volume contains thousands or even millions of duplicate data objects. As these objects are modified, distributed, backed up, and archived, duplicate data objects begin to proliferate at an alarming rate. The result is inefficient use of storage resources. NetApp deduplication helps prevent this inefficiency.

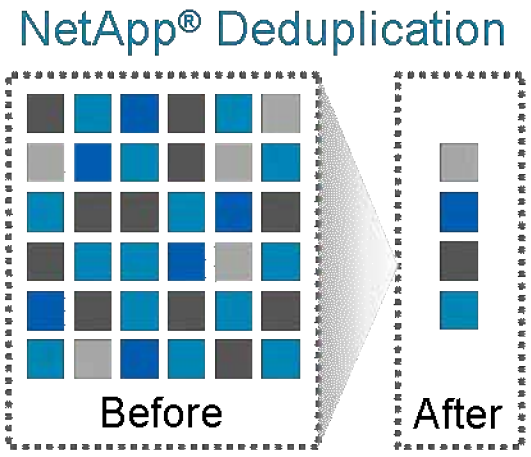

Figure 4) Deduplication removes data redundancies.

NetApp deduplication divides newly stored data objects into small blocks. Each block of data has a digital "signature," which is compared to all other signatures in the data volume. If an exact block match exists, the duplicate block is discarded and its disk space is reclaimed. Deduplication can be implemented across a wide variety of applications and file types, including data backup, data archiving, and unstructured data volumes. By implementing deduplication, customers can reclaim up to 95% of their storage space!

According to Tony Asaro, senior analyst at Enterprise Strategy Group, "Deduplication, at its core, is another form of data virtualization, in which one physical copy represents many logical copies. Deduplication creates a domino effect of efficiency, reducing capital, administrative, and facility costs. We believe that data deduplication is one of the most important and valuable technologies in storage."

USER CASE STUDY 3: FLEXVOL

NetApp is not only a producer but also a consumer of its storage arrays. In 2006, the NetApp IT group undertook a project to increase storage utilization. This project involved migrating from old, inefficiently used storage systems to new, more scalable systems using data virtualization techniques. This consolidation yielded significant results:

- Average storage capacity utilization increased from 40% to 60% per volume
- Storage footprint was reduced from 24.83 racks to 5.48 racks
- 50 storage systems were reduced to 10 storage systems
- Direct power consumption decreased by 41,184 kWh per month
- Annual electricity cost was reduced by $59,305

According to the director of facilities at NetApp, “NetApp is not only committed to providing customers with leading-edge data management techniques to reduce their data center power consumption, but also to decreasing our own energy usage with more energy-efficient technology. NetApp was facing several challenges, including growing space, cooling, and power constraints. Server and storage consolidation attacked the power consumption issue at the source.”

8 IMPROVING THE EFFICIENCY OF NON NETAPP STORAGE

As system administrators implement and realize the value of storage efficiency on their NetApp systems, a common comment is, “I wish I could apply this efficiency to my non NetApp storage too.” The NetApp V-Series gateway system allows just that; it is the first and only storage virtualization solution that unifies block and file (NAS, FCP SAN, and IP SAN) storage networking paradigms under a common architecture. With V-Series systems, your entire storage infrastructure can be virtualized under one interface. Storage arrays from HDS, HP, EMC, 3PAR, Fujitsu, IBM, and others are supported behind the V-Series controller. With a V-Series system, you can apply all the efficiencies described in this paper to your non NetApp systems.

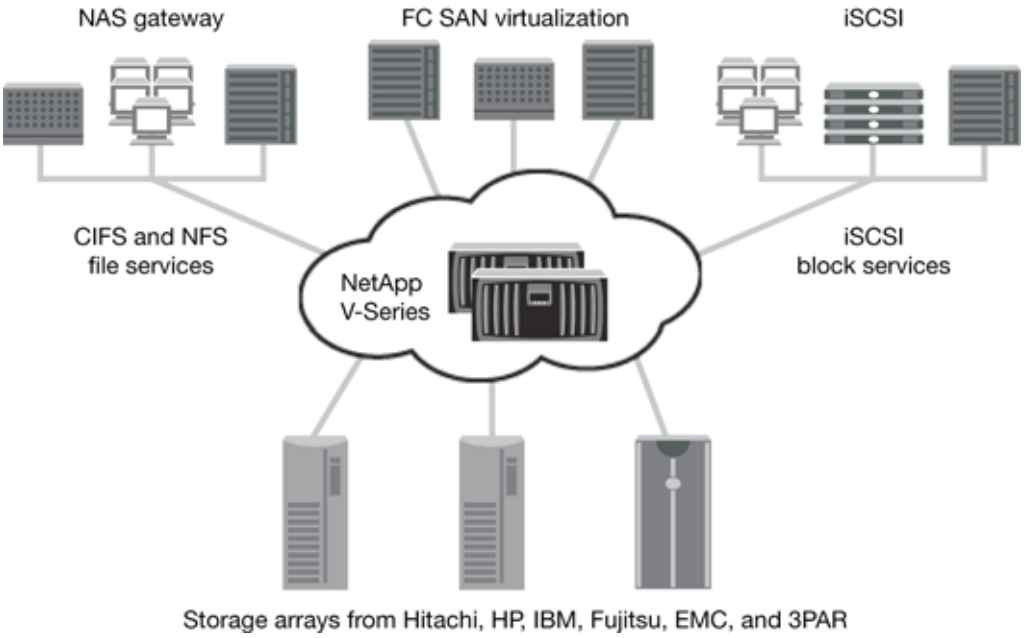

Figure 5) Using V-Series with non NetApp storage.

10 CONCLUSION

Data proliferation is a fact of life. Harnessing the spread of data is a near impossibility, but correcting provisioning inefficiencies; managing storage allocation for test and development, data backup, data replication; and eliminating data redundancies are well within the grasp of system administrators with today's storage efficiency tools. NetApp provides many products and features that enable space-efficient storage and result in **buying less storage**:

- **Snapshot:** Point-in-time file-system views
- **FlexVol:** Thin provisioning for quota efficiencies
- **FlexClone:** Writable test and development copies
- **SnapMirror:** Efficient data replication using Snapshot copies
- **SnapVault:** Disk-to-disk data backups with 20:1 or greater space-saving ratio versus tape backups
- **NetApp deduplication:** General-purpose deduplication for removal of redundant data objects
- **V-Series:** NetApp space efficiencies brought to non NetApp storage

As demonstrated in the user case studies, implementing NetApp storage efficiencies can have a significant impact on bottom-line expenses as they relate to the cost of storage acquisition as well as the ongoing costs of data storage management.

11 RESOURCES

The following related documents are also available from NetApp.

WHITE PAPERS

WP-7003-1106 "Maximizing Storage Utilization"

www.netapp.com/ftp/max-storage-utilization.pdf

WP-7017-0307 "NetApp Thin Provisioning: Better For Business"

www.netapp.com/ftp/wp-thin-provisioning.pdf

TECHNICAL REPORTS

TR-3509 "Power Consumption in the Data Center"

www.netapp.com/library/tr/3509.pdf

TR-3505 "NetApp Deduplication Deployment and Implementation Guide"

www.netapp.com/library/tr/3505.pdf