NETAPP WHITE PAPER

The NetApp Storage Efficiency Guide

Larry Freeman, NetApp, Inc.

July, 2011 | WP-7022-0711

EFFICIENCY TECHNIQUES TO ACHIEVE END-TO-END STORAGE CAPACITY REDUCTION

Table of Contents

1 EXECUTIVE SUMMARY

Several dynamics impact the way organizations approach IT purchasing decisions today. Lingering global financial uncertainty in the post-recession era has forced data center managers to re-evaluate their budgets and eliminate unnecessary IT spending. At the same time, environmental concerns continue to plague facilities managers, who are concerned with the cost and availability of power, cooling, and space required by today's growing 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, perhaps 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. The majority of system administrators are grappling today 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 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 systems, but all that data still resides somewhere, and is 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 new storage systems, and purchase smaller storage systems to begin with. You may not be able to control the behavior of your users or the pace of data growth, but you can control the efficiency of the storage systems that store this data.

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 budgets.

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 possible cost. NetApp Snapshot™ technology, introduced in 1992, was arguably the first widely-used storage efficiency feature available on disk-based enterprise storage systems. 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 have normally 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, 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 storage requirements by eliminating redundant data blocks.

3 MEASURING STORAGE EFFICIENCY

Just how do you measure storage efficiency? Believe it or not, this is a topic that has been debated for many years at NetApp. One view is that efficiency is anything that saves money or saves time. In the context of storage, using smaller systems to do the work of larger ones certainly saves money - but the time component of storage efficiency is a little less certain. Time to provision? Time to backup/restore? Time to read and write production data? Time to deploy applications? Putting a value on time and then trying to measure time efficiency is something we may eventually capture, but in this paper we will discuss storage efficiency in the context of reducing storage capacity requirements and saving money by using smaller storage systems.

Capacity Optimization is a good way to describe what NetApp does with storage capacity. When discussing optimization, it's helpful to break storage capacity into 5 categories:

1. **Raw Capacity**- This is a total addressable capacity of the storage system, before any setup is done
2. **System Reserve Capacity** - This is the capacity required during setup for things like system overhead, RAID parity and spare drives. This capacity is not available for user data
3. **Storable Capacity** - After the setup is complete - this is the total estimated capacity that *could* be used to store user data
4. **Stored Capacity -** Once user data is written, this is the amount of data that is currently stored on the system.
5. **Available to Store Capacity** - this is the estimated amount of *additional* user data that could be stored on the system before it fills up

Below is a graphical model that visualizes each of the above categories:

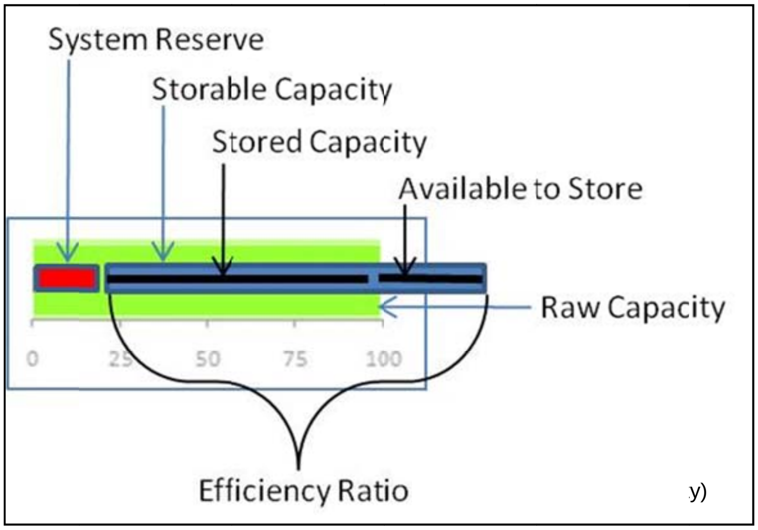


Figure 1) The Storage Efficiency Model.

Notice on the model above that the Storable Capacity bar extends to the right beyond the Raw Capacity bar. This is where capacity optimization is factored. Features like deduplication, compression, virtual clones, and snapshot copies make the storage system appear larger than it actually is. Storable Capacity is therefore an estimation of the effective amount of data that *could* be stored if efficiencies were enabled. Turn on the efficiencies, and the Storable Capacity bar moves to the right, turn off efficiencies, and Storable Capacity bar moves back to the left.

Storable Capacity is similar to the "Miles to Empty" indication used in many cars today. When you fill the tank, the display might indicate that you can drive 350 miles until the fuel tank is empty. That figure is an estimate based on average driving conditions for that car. If you drive with either a heavy foot or a light touch, your mileage to empty will end up being less or more than 350 miles. As you drive, the car's computer continuously updates and become more accurate as you approach zero miles, based on your driving habits.

Similarly, Storable Capacity is an estimate of the amount of data the storage system *could* hold based on average use of efficiencies. As data is stored on the system, calculations are updated based on the actual efficiencies in use and their effect on the data. As more and more data is stored, the Storable Capacity estimation becomes more and more accurate.

A reasonable goal for system administrators is to insure that all storage systems operate at 100% or greater efficiency, as the graphic model above demonstrates. 100% efficiency is attainable even with modest implementation of NetApp’s storage efficiency technologies, described in the remainder of this paper.

4 EFFICIENCY (OR INEFFICIENCY) 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 be a database entry, user file, software source code file, or email attachment, this data consumes physical space on a disk drive somewhere within your primary storage environment. Traditionally, the creation of data on primary storage begins a chain of events that lead to storage inefficiencies.

One of the first problems 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 (or LUN) 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 cannot 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 pattern of guesswork and resulting inefficiencies is not limited only to primary storage. Ineficiencies 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.

5 THIN PROVISIONING

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 and LUNs that appear to be a certain size but are in reality much smaller physically.

FlexVol technology provides substantial improvements in storage provisioning, utilization, and capacity 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.

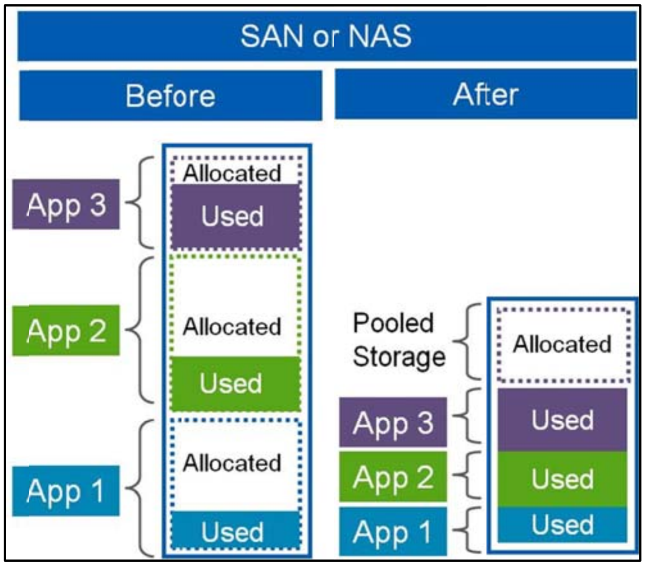


Figure 2) Thin Provisioning Allows for “Just-In-Time” Volume and LUN Provisioning.

6 VIRTUAL CLONING

Often, system administrators are required to allocate substantial primary storage capacity for essential enterprise test operations, such as bug fix testing, platform and upgrade checks, multiple simulations against large datasets, 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 the FlexClone^{TM} 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 dataset clones with minimal storage overhead. This is accomplished by creating a virtual copy of the primary dataset and storing only the data changes between a parent volume and a clone. All unchanged data remains on primary storage, and is utilized by both the primary application and the secondary clone copy. Multiple clone copies can be created from a single primary dataset, enabling users to perform multiple simulations and compare the characteristics of each dataset after the simulations are complete.

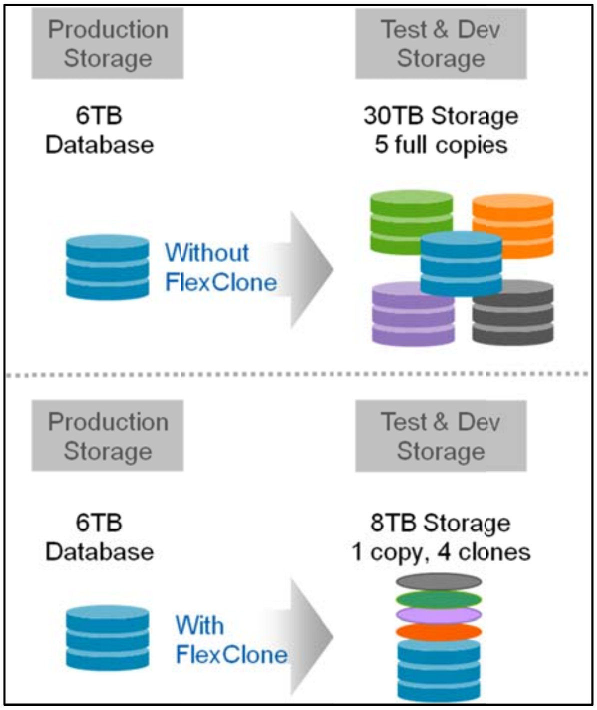


Figure 3) Using FlexClone to Create Virtual Test and Development Copies.

7 DATA REPLICATION AND BACKUP

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, 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 that last Snapshot copy, SnapMirror significantly reduces network bandwidth requirements. Data compression and deduplication, described later in this paper, facilitates further space reduction at the primary system, the secondary system, and during the data transfer between the two systems.   The result is lowered infrastructure cost of data replication and disaster recovery.

Similarly, SnapVault® protects data at the block level–copying only the data blocks that have changed since the last backup, not entire files. This means that you can take more frequent backups during the day. Storage capacity requirements are reduced because no redundant data is moved or stored.

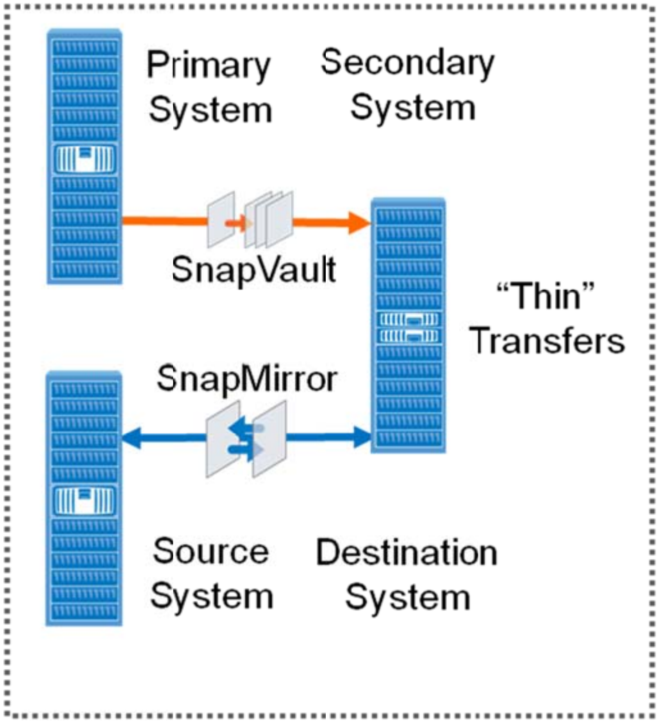


Figure 4) SnapVault and SnapMirror Enable Space-Efficient Backup and Disaster Recovery.

8 DEDUPLICATION

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 ax. 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 volumes grow ever larger. As mentioned in the opening section, 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 to prevent this inefficiency.

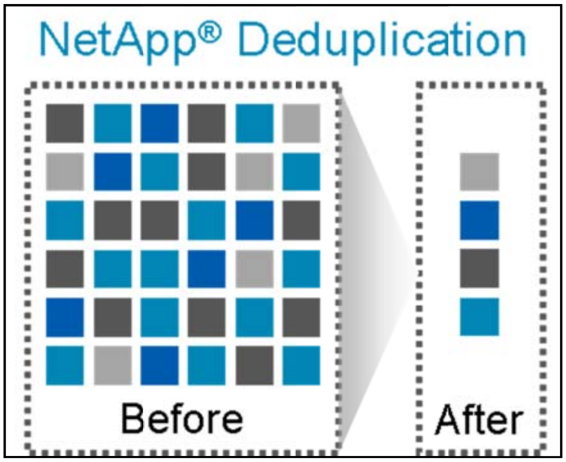
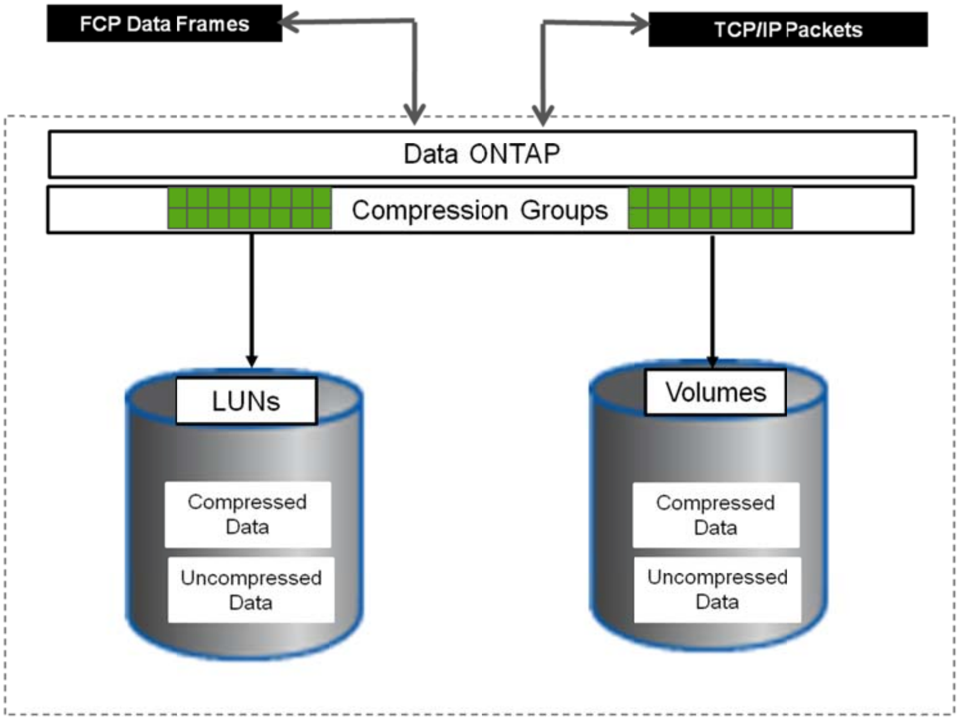


Figure 5) NetApp Deduplication Removes Data Redundancies

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!

9 DATA COMPRESSION

Data compression algorithms have been used for decades in tape devices and more recently in Virtual Tape Libraries. Data compression, while effective in reducing data capacity, exacts a high software performance penalty or requires the use of expensive hardware components – both of which has prevented widespread use on primary storage systems. NetApp data compression provides an optimal balance by using tried-and-true software compression with a twist. As data enters the storage system, it is fashioned into 32K compression groups. Each group is tested individually for compressibility. If a group is deemed compressible, it is stored in its compressed format; otherwise it is allowed to pass through in its original form. By compressing data selectively as needed, NetApp data compression offers substantial savings without a substantial performance penalty.



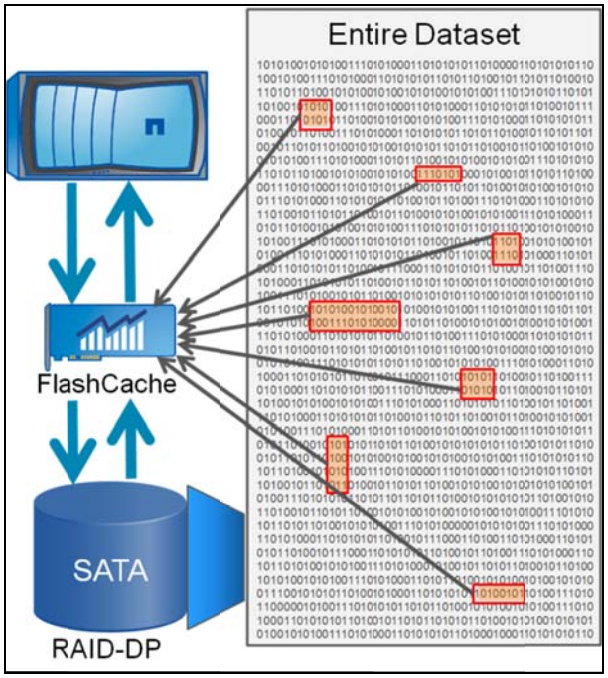
h

10 EFFICIENT USE OF HIGH CAPACITY DRIVES

High capacity SATA disk drives are the undisputed price leader in cost per GB; however, there are two concerns that have prevented them from being used for enterprise applications. First, SATA drives have a reputation of being less reliable than 15K rpm Fibre Channel or SAS drives. Secondly, SATA drives are slower than 15K drives because of slower rpm and actuator access rates.

NetApp addresses these concerns by combining SATA drives with two important technologies: RAID-DP and Flash Cache. RAID-DP provides additional device protection by tolerating dual drive failures without data loss – and without a significant performance penalty. Flash Cache automatically promotes “hot” data into a virtual storage tier, reducing the dependency on disk drive speeds by placing frequently accessed data into a fast solid-state memory cache.

By combining SATA drives with RAID-DP and Flash Cache, NetApp brings to market a new paradigm – high density drives used for high performance applications. The result is lower cost per GB as well as reductions in power, space and cooling requirements.



gh-

11 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 (FC, FCoE, iSCSI, CIFS, NFS) storage networking protocols under a common architecture. With V-Series systems, your entire storage infrastructure can be virtualized under one interface. Storage systems from HDS, HP, EMC, 3PAR, Fujitsu, IBM and others are supported behind the V-Series controller. With a V-Series system, you are able to apply all the software efficiencies described in this paper to your non-NetApp systems.

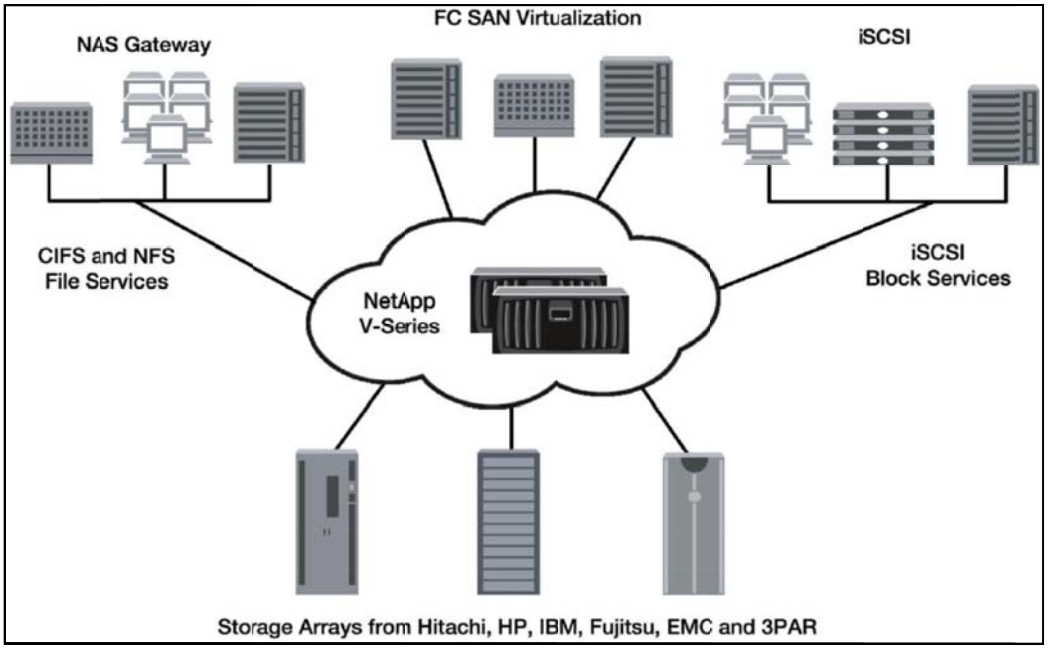


Figure 8) Using V-Series with Non-NetApp Storage.

12 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, resulting in **buying less storage** with NetApp:

- Snapshot: Point-in-time file-system views
- Thin provisioning: Just in time storage for provisioning efficiencies
- FlexClone: Writable test and development copies
- SnapMirror/SnapVault: Efficient data replication and backup using Snapshot technology
- NetApp deduplication: General-purpose deduplication for removal of redundant data objects
- NetApp Data Compression: Software data compression without significant performance degradation
- SATA/RAID-DP/Flash Cache: Use high density storage for high performance applications
- V-Series: NetApp space efficiencies brought to non-NetApp storage

As demonstrated in the following 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.

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%.”

USER CASE STUDY 2: DEDUPLICATION

A major multimedia company is also a long-time NetApp customer. Among other applications, this company has 3 SQL servers with a total capacity of 2TB. These databases are considered essential to operations, containing customer billing information. From the main data center, all three SQL databases are backed up nightly to a NetApp storage system in a second location. From that location, the three databases are again backed up to a third location for disaster recovery (DR) 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 NetApp 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 result of this implementation was a completely automated database backup process, and a 40% reduction in disk requirements, reducing the capacity requirement from 32TB to 19TB.

USER CASE STUDY 3: UTILIZATION IMPROVEMENT

NetApp is not only a producer but also a consumer of its storage systems. The NetApp IT group recently undertook a project to increase storage utilization. This project involved migrating from old, inefficiently used storage systems to new, more scalable systems utilizing 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.”