HITACHI
Inspire the Next

DATA DRIVEN GLOBAL VISION CLOUD PLATFORM STRATEG
ON POWERFUL RELEVANT PERFORMANCE SOLUTION CLO
VIRTUAL BIG DATA SOLUTION ROI FLEXIBLE DATA DRIVEN V

WHITE PAPER

Hitachi Unified Storage: Break Down Capacity Efficiency Barriers on Primary Storage Through Deduplication

By Hitachi Data Systems

September 2013

Hitachi Data Systems

Contents

Hitachi Unified Storage: Break Down Capacity Efficiency Barriers on Primary Storage Through Deduplication

Executive Summary

As organizations endeavor to innovate and find competitive advantage, the need for a highly nimble and cost-efficient data center has never been greater. Organizations of all sizes are challenged by unprecedented data growth demands and the costs and complexities that follow, especially with flat or shrinking budgets.

The traditionally separate block and file storage environments are not equipped to support the wisdom or realities of today's new data requirements. Greater emphasis is placed on deriving fresh business value from data and addressing the proliferation of unstructured content.

The savvy storage shop is looking to unified storage platforms for answers to today's most difficult data growth imperatives. When block and file data are united, IT can more quickly increase storage capacity efficiencies, including improved storage utilization and data accessibility. Bottom line, IT departments can deliver the necessary storage services to their customers at the right cost.

Hitachi Unified Storage (HUS) offers an intelligent, automated and performance-centric solution capable of meeting immediate and long-term requirements. The advanced capacity efficiency technologies in HUS allow IT leaders to control unruly data storage growth while substantially lowering the total cost of ownership (TCO).

This white paper discusses the capacity efficiency technologies in Hitachi Unified Storage, with particular focus on the primary storage deduplication capabilities included in HUS and in Hitachi NAS (HNAS) Platform. Hitachi primary storage deduplication helps deliver up to 90% improvement in capacity efficiency without sacrificing performance, simplicity or scalability.

Introduction

When it comes to overcoming data storage challenges, nothing has changed — and everything has changed. Organizations are still juggling heavy burdens, such as trying to solve complex storage issues while attending to flatlined IT budgets and greater demands. What is markedly different are the immense volumes and types of data being generated and stored, and how to repurpose it for greater business value.

No question: Data centers must become more efficient and simpler to adeptly meet the myriad of challenges for the impending future. For instance, the rate of data growth is unparalleled. While technology is continually renovating the ways business operates, so is the perpetual expanse of the digital universe. A recent data point from analyst group IDC[1] forecasts a 5-fold growth factor in this decade of the amount of digital content created by individuals and businesses, reaching 40 zettabytes (ZB) by 2020. To put this in context, 40ZB is equal to 57 times the number of every grain of sand on all beaches on earth.

Data Growth Challenges

Some of the most urgent data growth challenges facing IT today include:

Unyielding Growth of Unstructured Content

It should come as no surprise that the overwhelming majority of incoming data is unstructured content. Workforces are much more mobile, accessing enterprise applications from smartphones and wireless devices, in addition to storing the same files on office desktops, laptops and servers. Organizations in industries such as healthcare and manufacturing are deploying virtual desktop infrastructures to quickly access vital information from anywhere, at any time. Rich content files can range from e-discovery materials and media files to computer-generated renderings and medical imaging records. Web-based content and the growing use of digital video and radio-frequency tags can quickly add bulk, as well.

Need for Greater Business Value

Alongside the ever-increasing amounts of storage to manage is the morphing value of the data. The acute focus now is on how to derive greater value from existing and inbound data. Evolving data into meaningful information can provide organizations with greater business insights and competitive advantage.

Data is changing how organizations do most everything. Eliciting more intelligence from stored data is essential to leveraging information as a strategic asset. Many enterprises actually have unknown amounts of untapped data, lurking across siloed systems. When that data becomes unified and readily accessible, the organization can tap a new source of information for improving decision-making, accelerating time to market, or improving the customer experience.

Reduce Storage Costs

As businesses strive to gain momentum in a vacillating global economy, the ability to cost-effectively solve complicated data management issues has never been greater. IT leaders are paying particular attention to opportunities to lower the TCO and reach a return on investment (ROI) more quickly.

Storage is usually the largest single line item of an enterprise IT capital expenditure (capex) budget. Yet, the TCO goes beyond the price of hardware to encompass operational expenditures (opex), much of which relates to the management of storage. Typically, as the storage environment surges in size and complexity, so does the cost to manage it.

[1] IDC Digital Universe 2012

Midsize to enterprise, businesses are looking to control costs and simplify the management of valuable information. Emphasis is on technologies that can alleviate cost inefficiencies and reduce the strain on resources.

IT Administration Complexities

The list of administrative demands is long: Maintain or exceed service levels for application users and customers. Improve performance and reliability across the data center. Provide faster application deployment and access to stored data. Add new services. Meet service level objectives and rising customer expectations. In other words, do more with less, and do it better.

Administrators are continually seeking solutions that can reduce complexity, risks and costs. Executives are looking to deliver storage services on time and under budget. Conventional methods and technologies were not built to accommodate the exponential growth rates of data storage. IT departments are no longer forecasting terabytes of growth but petabytes and beyond. Managing data at the petabyte scale requires a mega-efficient model that is sustainable and promises greater proficiency and simplicity at multiple levels.

Barriers to Capacity Efficiency

Capacity efficiency is a mantra for many organizations looking to simplify infrastructure, extend the value of storage assets, and strip out unnecessary expense. Capacity is the physical amount of storage available on any given system. Capacity efficiency is determined by how well administrators are able to use system storage to benefit the entire storage environment. Two common indicators for capacity efficiency are storage utilization and availability.

Siloes and Storage Utilization

Status quo solutions, such as traditional siloed file and block storage environments, no longer measure up to imperative capacity efficiency goals. While NAS and SAN environments commonly co-exist in many medium-to-large-scale data centers, they have not necessarily worked as unified storage. When storage is segregated or leads to unintended silo sprawl, the results are often poor utilization rates and higher costs across the entire storage landscape.

Data inefficiencies can grow faster than the data itself. An organization experiencing cumulative data growth across silos is compounding the complications and overhead of storing it. Finding ways to more easily connect and reclaim unused capacity is essential to improving utilization rates, service levels and cost structures.

An important industry trend is finding unified storage solutions that offer all protocols. For file data, that means SMB, NFS, FTP and MPFS protocols. Block data is accessible through both Fibre Channel and Fibre Channel over Ethernet (FCoE), or by iSCSI protocols. By bringing together all data under a unified umbrella, the confines of typical data storage give way to greater flexibility and cost savings.

Storage Availability Across Systems

Administrators need more than just raw capacity. They also need storage strategies that help to prioritize data and provide a high standard of service. Availability of storage capacity is vital to ensuring access to critical data and business applications. When data is more effectively accessible, the organization will enhance productivity of systems and operations, and can better promote business goals.

Typically, when storage environments are not unified, storage hardware can sit idle or may not be accessible or freely available. An inability to leverage existing disk space directly affects the storage manager’s ability to free up capacity, resources and costs.

When storage systems are not unified, there is an inherent complexity; there is a lack of visibility to see what capacity exists across systems and limited dexterity to manage it. Organizations need automated, at-a-glance unified capabilities to readily access and control stored data, regardless of where or on which systems it resides.

Capitalize on Storage Capacity Efficiency Practices

Attaining capacity efficiencies of noteworthy significance can be likened to earning a merit badge for thriftiness. The attention is on using practices and technologies to improve storage utilization, reclaim capacity, repurpose existing systems, and take advantage of cost-saving features.

Consolidation and Simplification

Production storage environments nearly always have opportunities to consolidate errant storage. Consolidation frees up high-performance storage capacity for future use and helps extend the useful life of existing assets. In turn, IT can begin to reclaim valuable storage capacity and potentially defer related purchases. The future purchase of capacity can now be more in line with business growth rather than dictated by the storage environment's inefficiencies.

Reducing the quantity of storage systems through consolidation efforts will minimize energy costs and the number of floor tiles needed to house less equipment. Consolidated storage also means an opportunity to reduce administrative time spent managing and maintaining systems.

Unified File and Block Storage

The value of capacity efficiency goes way up when consolidated storage systems are unified as a single pool to manage. Storage systems operating as a single entity rather than scattered resources enable organizations to take advantage of cost-saving features. Technologies that can be shared across the storage pool help to reduce data redundancy and management practices, and diminish time-consuming manual tasks. Even events such as migration and backup windows can be abridged and simplified. Unifying file and block storage allows for a cohesive and more accurate view of storage requirements, performance, utilization and growth forecasts.

The Technologies to Get There

The most prominent technologies that drive up capacity efficiency are thin provisioning, tiering, migration and capacity optimization. In addition, storage virtualization can extend these capacity efficiency technologies to enable customers to reclaim capacity and increase utilization and visibility across a larger range of storage assets.

Thin Provisioning

In traditional storage provisioning, large amounts of physical capacity are pre-allocated in anticipation of an application's needs, and storage capacity is consumed immediately whether data is written to it or not. In contrast, thin provisioning allows physical capacity to be allocated from the virtualized pool when it is needed. Disk capacity consumption is delayed until the application actually writes data to that capacity. Thin provisioning is a mature and stable technology, which can increase storage utilization rates by as much as 60% to 80%.

Automated Tiering

Increased microprocessor performance within storage controllers facilitates the implementation of more complete tiering algorithms that can track real-time data access patterns. Automated tiering algorithms can move data, detect patterns and make decisions 24 hours a day in real time and can really improve application performance for storage administrators. Tiering up for performance with faster disks or solid-state is only half of the use case. By tiering down, IT can realize cost savings. Tiering down from solid-state to SAS or SATA when hot data turns colder is just as valuable, if not more so, than tiering up.

Data Migration

Data migration moves data automatically between different storage tiers based on user-defined policies. Data can be demoted or promoted as performance and protection requirements of the business change at any point in time.

Caching allows a background copy to be available for subsequent requestors of a file. If we think of migration of data for economic tiering reasons, caching is more for performance tiering. Dynamic read caching permits the files of unstructured data sets to be instantly copied to a higher performance storage tier. When data migration and caching are combined, the result is extensive flexibility.

Data Reduction

Capacity optimization is accomplished through compression and data deduplication. These are complementary data reduction technologies, which serve to reduce costs of ownership for storage infrastructure. Compression works to remove unnecessary byte patterns (small bits of data) from a file or object, whereas deduplication eliminates redundant data to allow administrators to actually store more data on fewer disks.

The remainder of this white paper will focus on the capacity optimization provided by data deduplication.

Data Deduplication Basics

While data deduplication offers prospects for impressive capacity efficiency, the seasoned IT leader is all too familiar with the broken promises of many dedupe solutions. Perhaps the assurance of increased capacity was broken with limitations and restrictions that delivered only minimal savings on storage space. Or a product pledge to dedupe without impact to performance in fact brought down the overall system performance. The declaration of a solution that will deduplicate everything stored actually only addresses a portion of total capacity and is limited to terabytes. Anticipation of data deduplication solutions that promise to do away with wasted space, run automatically or minimize management have all been frequently met with disappointing realities.

So, is there a redeeming set of promises that data deduplication technologies can deliver? Let’s further examine how deduplication works, the types of deduplication technologies available and how they are poised to dispatch with capacities inefficiencies.

File Versus Block-Level Deduplication

Data deduplication works at either the file level or block level of storage to eliminate redundant data. File-level data deduplication starts by retaining only 1 unique instance (or copy) of the data on storage media, while duplicates are replaced with pointers to this single copy. A comprehensive index is maintained so that all data can be transparently accessed.

To illustrate the effect of file-level deduplication, let’s use an email example. If an email with a 1MB file attachment was sent to 100 users in the same organization, the email system might have 100 instances of that same file attachment. This would require 100MB of storage space. With file deduplication, since only 1 instance of the attachment is stored, there is an immediate occasion to dramatically control data size and jettison unnecessary copies for any given piece of content. In our example, data deduplication would eliminate all but the original 1MB, allowing the other 99MB to be removed as redundant. Large gains can be realized with eliminating redundant files from unstructured data.

Perhaps larger efficiency gains can be seen with block-level deduplication. This type of deduplication is more granular and operates at the volume level by deduping the blocks of data that make up that volume. For example, let’s say there are 3 versions of a document, each just a little different than the others. With block-level deduplication, the 1st document, which is unique, will be stored in its entirety. For all other versions of that document, only the changed portion of the document will be stored. In contrast, with file-level deduplication, all 3 versions of the document would be stored in its entirety because each is slightly different and thus unique.

Both types of data deduplication offer additional pros and cons, but in general, block-level deduplication can provide better capacity optimization.

Primary Versus Secondary Deduplication

Deduplication technologies can be applied to primary storage and secondary storage. Primary storage is defined as containing active, online or near active and transaction data, all of which are highly accessible to end users. The primary storage systems therefore are designed for optimal performance. Secondary storage is used to house less active data, such as historical content, archival or backup data.

To date, deduplication has been chiefly used with secondary storage systems. The main reason is that archival and backup applications generate substantial amounts of duplicate data over time. The primary goal of deduplication on secondary storage is to store this massive amount of data on the smallest amount of disk to improve efficiency and reduce costs.

On the other hand, primary storage systems mandate superior performance, even at the expense of other considerations, and are less tolerant of any operation that would adversely affect performance. Most deduplication technologies have been seen as creating overhead that would impact performance. And, while the presumption that duplicate data exists is pretty certain for secondary storage, IT leaders traditionally assume the opposite for primary storage: that it seldom contains duplicate data. For these reasons, IT has conventionally sidestepped data deduplication practices here.

Recent research indicates, in fact, that up to an average of 40% of the data in primary storage may indeed be duplicate data. For example, copies of data created for testing a new application on primary storage were perhaps never removed post testing. In virtualized environments, the amount of redundant data soars and can reach 90% or higher.

Only a couple of years ago, primary dedupe was not considered viable as a mainstream strategy for improving production environment storage. The garden-variety storage optimization tools are not engineered to manage the high transactional traffic and random I/O concentration of many production environments.

At present, primary storage deduplication is steadily gaining a foothold as the costs of running production environments continue to escalate. This pushes IT organizations to operate precious primary storage assets at optimal levels to meet critical business objectives and lower capex and opex costs.

The concern about performance impacts on primary storage systems may have dampened the perception of the power of data deduplication. Performance degradations would make dedupe impractical to use during primetime file serving. Finding data deduplication technologies that do not compromise performance across the storage environment is basic to capacity efficiency goals.

Enter Hitachi, with primary storage deduplication that delivers on the promises of deduplication without compromising performance.

Hitachi Delivers Capacity Efficiency Without Compromise

Hitachi Unified Storage unifies all data types so organizations can capitalize on business benefits in cost-effective ways. With best-in-class scalability, performance and availability, Unified Storage optimizes support for critical applications, cloud-ready infrastructure, and data center consolidations, all through a single intuitive interface. File data is accessed on Unified Storage through SMB, NFS, FTP or MPFS protocols, while block data can be accessed via Fibre Channel, FCoE or iSCSI protocols.

Hitachi Unified Storage is not just another block and file product. Instead, it delivers on the Hitachi Data Systems vision of an integrated portfolio of intelligent storage solutions that unify the storage environment for high reliability and increased capacity efficiency through integrated tiering, migration, thin provisioning and optimization technologies.

Primary Storage Deduplication That Delivers

Hitachi Data Systems has introduced primary storage deduplication that truly delivers on the broken promises of other storage vendors in the past and brings capacity efficiency and business value without compromise. The deduplication capabilities found in the Hitachi NAS Platform and HUS carry some heavyweight wins for meeting enterprise-level demands with enterprise-level functionality (see Table 1). They leverage the hardware-accelerated architecture (FPGA), which provides high-performance file services to execute data deduplication.

The Hitachi NAS Platform architecture includes an object-based file system offload engine (FOE) that is powered by powerful integrated FPGAs. The FOE handles hashing and chunking, the most processor-intensive tasks of deduplication, via the accelerated hardware, rather than through software, to ensure performance is protected. In fact, it uses up to 4 engines working in parallel to dramatically increase deduplication performance nearly 4-fold.

TABLE 1. MEET ENTERPRISE-LEVEL DEMANDS WITH ENTERPRISE-LEVEL FUNCTIONALITY.

The Promise of Hitachi Primary Storage Deduplication

The Promise	Promise Kept?	The Reality With Hitachi NAS Platform
Increased capacity	Yes!	Up to 90% capacity savings
No performance impact	Yes!	Automatically throttles back when busy
Dedupes everything	Yes!	Entire capacity can be deduplicated
No wasted space	Yes!	Minimal overhead = maximum efficiency
Runs automatically	Yes!	Automatically starts for any new data
Minimal management	Yes!	Set it and forget it!

Some of the other key features of Hitachi primary storage deduplication include:

Auto Throttling Deduplication

Timing is everything. With up to 4 high-speed deduplication engines, redundant data is automatically eliminated when the system is not busy. When file serving loads reach 50% of the available IOPS, the dedupe engines throttle back. This transparent “throttle gatekeeper” prevents any impact to user performance when the system is busy and then automatically resumes when the system is less busy.

Automated Operation

Hitachi uses an intelligent “set-it-and-forget-it” deduplication process, which is aware when new data is added and automatically starts up the deduplication engines. In this way, administrators are able to eliminate the need to spend valuable time configuring, tuning or scheduling deduplication operations.

Data-in-Place Deduplication

With Hitachi, data is stored as it normally would be. The unique data-in-place deduplication technology combs through data looking for redundancy and eliminates duplicates where they are stored. Other products on the market require a significant portion of pre-allocated disk to be set aside as temporary dedupe workspace, which is wasteful. It is also contrary to the primary objective of deduplication, which is to improve capacity utilization.

Enterprise Scalability

Unlike other NAS solutions on the market that severely limit deduplication to only a small portion of overall storage capacity, the Hitachi patent-pending deduplication algorithm allows for the entire usable storage capacity to be deduplicated. HNAS also includes a multipetabyte global namespace to virtualize several file systems, even deduplicated ones, under a common addressable space. This ensures enterprise scalability so organizations can quickly realize extensive savings through capacity purchase avoidance or deferral.

Minimal Overhead

When usable storage capacity is prized, why would IT want to submit to any unnecessary overhead requirements? Unlike other solutions that waste 7% or more of usable capacity, the Hitachi solution requires only a miniscule amount of disk to track deduplicated data.

Nondisruptive Implementation

Hitachi makes life simpler for storage administrators with its optional deduplication feature, which can be easily implemented on new or existing systems without making significant changes to existing computing environments.

Cost Benefits of Hitachi Deduplication

The net benefit of Hitachi primary storage deduplication is lower TCO, which is achievable through increased capacity efficiency without sacrificing performance and scalability. The set-it-and-forget-it nimbleness means less manual intervention and streamlined administrative processes, and, more importantly, reductions in complications, error, risk and costs.

Lower Capex

Hitachi primary deduplication eliminates or defers the need to purchase more capacity or additional systems to meet increasing storage demands. IT does away with the commonplace need for a temporary dedupe workspace, thereby minimizing the space needed to orchestrate data deduplication most efficiently.

Lower Opex

A reduction in operating expenditures is realized with fewer systems and less capacity to manage and maintain. There should also be a reduction in required licensing, power and cooling costs, and floor space. Intelligent “administration-less” deduplication, with no manual scheduling, configuring, tuning or monitoring, leads to savings in time spent managing storage.

CUSTOMER SUCCESS WITH HITACHI DEDUPLICATION

A Fortune 500 global semiconductor and electronics company is an early customer who evaluated Hitachi data deduplication capabilities alongside products from other vendors. The customer found Hitachi dedupe to be the most effective method, and was impressed with the speed of deduplication, using only the single deduplication engine included at no cost with the base system. The single engine deduped over 1.2 million files in only 16 minutes. The customer was further impressed by the fact that there was minimal impact to primetime file serving activity, and has chosen to deploy Hitachi deduplication in its environment.

Capacity Efficiency Beyond Deduplication

Hitachi Unified Storage from Hitachi Data Systems is a comprehensive platform for improving capacity efficiency beyond the extensive benefits of capacity optimization provided by primary storage deduplication. High-end storage functionality enables all file, block and object data to be provisioned, managed and archived throughout its lifecycle.

Hitachi Dynamic Tiering

Hitachi Dynamic Tiering simplifies storage administration, solves the complexities of implementing data lifecycle management and optimizes the use of tiered storage by eliminating the need for time-consuming manual data classification and data movement operations. Hitachi Dynamic Tiering automatically moves fine-grain data within these tiers, according to workload, to the most appropriate media. And, the more automation, the better. Built-in automation of data movement between tiers shrinks the time spent managing storage to increase productivity and reduce administrative costs.

Hitachi Dynamic Provisioning

Hitachi Dynamic Provisioning (HDP) enables physical capacity to be allocated on a just-in-time basis in 1GB chunks. HDP pools can expand and contract nondisruptively, which can drastically improve capacity utilization rates. Another benefit of HDP is its ability to distribute application workloads across many RAID groups, avoiding hot spots and improving performance.

Hitachi Content Platform

Hitachi Content Platform (HCP) is an object storage solution that allows IT organizations to store, share, synchronize, protect, preserve, analyze and retrieve file data from a single system. The system is more efficient, easier to use, and handles much more data than traditional archive solutions. HCP automates day-to-day IT operations such as data protection and readily evolves to changes in scale, scope, applications, and storage and server technologies over the life of data. In IT environments where data grows quickly or must remain for years, these capabilities are invaluable.

Conclusion

Hitachi Unified Storage and the Hitachi NAS Platform empower administrators to take back control over unruly data storage growth, with powerful capacity-optimization features such as primary storage deduplication. Hitachi deduplication technology helps to deliver up to 90% improvement in coveted capacity efficiency, without compromising on performance, usability or scalability. Hitachi has delivered on the promise of “deduplication without compromise,” enabling the benefits of increased capacity efficiency and lower TCO.

The Hitachi Data Systems File and Content Family

Hitachi Unified Storage is tightly aligned with the portfolio of Hitachi file and content solutions. With such attention paid to the state of unstructured data growth, the file and content solutions operate as an agile, integrated family of products designed to reduce cost and complexity, with:

- Support for cloud-ready initiatives.
- Intelligent policy-based tiering to lower cost media or other targets.
- Robust search and discovery tools.
- Advanced data protection.
- Support for multiple protocols, industry standards and applications.
- Strategic framework to archive first, back up less, and consolidate more.

Appendix A: References

Read the Latest Expert Blogs:

- Primary Storage Deduplication without Compromise
  By Hu Yoshida, Vice President and Chief Technology Officer
  Hitachi Data Systems
- The Storage Olympics Gets Magical
  By Bob Madaio on April 1, 2013
- Deduping Dedupe
  By Michael Hay on April 8, 2013

Where to Learn More:

- Hitachi File and Content Solutions Web page
- Hitachi NAS Platform: System Software v11.1 Datasheet
- Hitachi Unified Storage 100 Family Web page
- Hitachi NAS Platform Web page

Hitachi Data Systems

Corporate Headquarters
2845 Lafayette Street
Santa Clara, CA 96050-2639 USA
www.HDS.com   community.HDS.com

Regional Contact Information
Americas: +1 408 970 1000 or info@hds.com
Europe, Middle East and Africa: +44 (0) 1753 618000 or info.emea@hds.com
Asia Pacific: +852 3189 7900 or hds.marketing.apac@hds.com

Notice: This document is for informational purposes only, and does not set forth any warranty, expressed or implied, concerning any equipment or service offered or to be offered by Hitachi Data Systems Corporation.