HITACHI
Inspire the Next

Performance Best Practices Guide for a HNAS Solution – v3.0

A White Paper

By Gokula Rangarajan (File Services Competency Center - Technical Operations)

HDS employees and TrueNorth™ Partners only. NDA required for customers.

July 2013

Purpose of this paper

The purpose of this paper is to outline the performance related best practices and guidelines one should consider before advising, architecting and deploying a HNAS solution with high performance requirements. This paper outlines the factors affecting or contributing to the performance from both storage and the HNAS perspective. Some of the concepts and best practices noted in this paper are supported with performance results when available.

An HNAS environment typically consists of one or more high performance HNAS systems with LUNs from Hitachi midrange or enterprise class storage systems. These NAS systems are also referred to as 'nodes' "heads" "servers", and may be operated individually or as cluster pairs. In a standard HNAS environment, the storage array types include the Hitachi Unified Storage (HUS) 100 family or the Hitachi Unified Storage VM (HUS VM) or the Virtual Storage Platform (VSP) or the previous generation storage systems.

This document is designed for individuals who are responsible for advising, planning, designing, deploying and supporting an HNAS environment, with a goal of achieving the highest level of performance and availability. In order to design and deploy the right configuration, one should have a clear understanding of the customer environment, workload and expectations.

Notices and Disclaimer

The performance data contained herein was obtained in a controlled isolated environment. Actual results that may be obtained in other operating environments may vary significantly. While Hitachi Data Systems Corporation has reviewed each item for accuracy in a specific situation, there is no guarantee that the same results can be obtained elsewhere.

All designs, specifications, statements, information and recommendations (collectively, "designs") in this manual are presented "AS IS," with all faults. Hitachi Data Systems Corporation and its suppliers disclaim all warranties, including without limitation, the warranty of merchantability, fitness for a particular purpose and non-infringement or arising from a course of dealing, usage or trade practice. In no event shall Hitachi Data Systems Corporation or its suppliers be liable for any indirect, special, consequential or incidental damages, including without limitation, lost profit or loss or damage to data arising out of the use or inability to use the designs, even if Hitachi Data Systems Corporation or its suppliers have been advised of the possibility of such damages.

Hitachi Unified Storage®, Hitachi Unified Storage VM®, Hitachi NAS Platform® and Virtual Storage Platform® are registered trademarks of Hitachi Data Systems, Inc. in the United States, other countries, or both.

Other company, product or service names may be trademarks or service marks of others.

This document has been reviewed for accuracy as of the date of initial publication. Hitachi Data Systems Corporation may make improvements and/or changes in product and/or programs at any time without notice. No part of this document may be reproduced or transmitted without written approval from Hitachi Data Systems Corporation.

Document Revision Level

Revision	Date	Description
1.0	October 2011	Initial Release
1.1	November 2011	VSP SPC-1, sdpath and few other changes
2.0	May 2012	HUS 100, Superflush enhancements, Performance Accelerator Feature, Sector cache etc and few other changes
3.0	July 2013	HDP, HDT, HUS 100 new IORL limits, HUS VM, HNAS 4000 series, Chunk size, FMD and several changes.

Reference

- AMS 2000 Architecture and Concepts Guide
- HUS 100 Architecture and Concepts Guide
- VSP Architecture and Concepts Guide
- Architecture and Concepts Guide - HUS File Module
- Architecture and Concepts Guide - HNAS 4000 Family
- Several HNAS Performance briefs and presentations

Contributors

The information included in this document represents the expertise, feedback, and suggestions of a number of skilled practitioners. The author would like to recognize and sincerely thank the following contributors of this document (listed alphabetically):

- Alan Benway, Performance Measurement Group - Technical Operations
- Amit Chakraborty, File Services Competency Center - Technical Operations
- Chayan Sarkar, File Services Competency Center - Technical Operations
- Shekhar Berry, File Services Competency Center - Technical Operations

Reviewers

The author would also like to recognize and sincerely thank the following reviewers of this document (listed alphabetically):

- Alan Benway, Performance Measurement Group, Technical Operations
- Al Hagopian, Master Solutions Architect
- Andreas Krause, HNAS Technical Product Manager
- Jeffrey Blomberg, Technical Operations
- Thomas Rivera, Senior Technical Associate – File, Content & Cloud Solutions

Table of Contents

Storage Concepts & Best Practices

Before recommending a storage design for customers, it is important to know how the product will address the customer's specific business needs. The factors that must be taken into consideration include: Capacity, Performance, Reliability, Features and Cost with respect to the storage infrastructure component. The more data you have, the easier it is to architect a solution as opposed to just selling another storage unit. The types of storage configured, including the disk types, the number of RAID Groups, RAID levels, the number of host paths, etc, are all highly important to the solution.

Choose the correct Storage System

The **Hitachi Unified Storage (HUS)** 110, 130, and 150 models are the replacements for the previous AMS 2000 generation of midrange Hitachi modular storage systems. The new HUS 100 family systems have significantly higher performance than AMS 2000 and incorporate several significant design changes.

The HUS 110 storage system consists of two controllers, 8GB of cache, and three choices of host port types (8Gbps FC, 1Gbps iSCSI, and 10Gbps iSCSI) with up to twelve ports total. At the core of each controller there is a custom Hitachi DCTL processor and an Intel Core-i Xeon single-core processor.

The HUS 130 consists of two controllers, 16GB of cache, and two choices of host port types (8Gbps FC and 10Gbps iSCSI) with up to sixteen ports total. At the core of each controller there is a Hitachi DCTL processor and an Intel Core-i Xeon dual-core (X and Y) processor.

The HUS 150 consists of two controllers, 32GB of cache, and two choices of host port types (8Gbps FC or 10Gbps iSCSI) with up to sixteen ports total. At the core of each 150 controller there is a Hitachi DCTL processor and an Intel Core-i Xeon dual-core (X and Y) processor.

The **Hitachi Unified Storage VM** system fits in between the midrange HUS 100 family and the high end VSP model. The HUS VM offers much higher performance, scalability and reliability than any competitive offering today. HUS VM supports up to 1152 disks and delivers up to **181,492 SPC-1 IOPS.** Depending on the IO sizes and read/write distribution for sequential workloads, the HUS VM also delivers between **5** to **12.5 GB/s** throughput.

The **Virtual Storage Platform** (VSP) is Hitachi enterprise subsystem that supports a large number of intermixed SSD, SAS, and SATA disks. The VSP's upper limit on disks in a dual chassis configuration is 1280 (3.5" LFF) or 2048 (2.5" SFF) disks. The VSP delivers up to **269,507 SPC-1 IOPS** and depending on the IO sizes and read/write distribution for sequential workloads, the VSP delivers between **6.4 to 12.5 GB/s** throughput, versus **5.8 to 11.5 GB/s** on the USP V.

Note: The performance characteristics noted above are from the backend storage system perspective only. Refer to "Choose the correct HNAS system" section for HNAS specific performance characteristics.

Choose the correct disk drive type

The HDS storage systems currently support 15k RPM, 10K RPM, 7.2K RPM SAS disks (also referred to as NL-SAS), SATA, FMD and SSD drives. The characteristics of these drives are shown below in Table 1:

Table 1: High level Characteristics of Disk Drives

Characteristics	SATA	7KSAS (NL-SAS)	10K SAS	15K SAS	SSD/FMD
Cost per Drive	Low <----------				----------> High
Performance per Drive	Low <----------				----------> High

Every spinning disk drive has a maximum random small block IOPS rate determined by seek times (head movement across the disk) and rotational delays (disk RPM). The number and size of the LUNs created per RAID Group determine how much of a disk's surface is in active use. As more of the disk surface is allocated to LUNs in a RAID Group, the disk heads must seek farther for some requests and this creates higher average seek times. Tables 2 below illustrate the estimated IOPS per drive at various cylinder utilizations for random write workloads.

Table 2: Estimated peak IOPS limits for write workloads, broken down by % cylinder utilization

Drive Type	25% cylinder utilization	50% cylinder utilization	75% cylinder utilization	100% cylinder utilization
15K RPM FC Drive	244	202	187	161
10K RPM FC Drive	175	147	137	119
7.2K RPM SATA Drive	118	98	91	78

Note: Above are the estimated IOPS rate per disk drive when operating at 100% busy rate, Random workloads, 8KB I/O and for 100% Write workloads

Table 3 below illustrates the "rule of thumb" typical random 8KB random read IOPS rate for each type of disk drive.

Table 3: Typical maximum IOPS limits of various drives

Drive Type	Typical Physical Maximum Read IOPS per Drive
SAS 3.5" 7.2K RPM	80
SAS 2.5" 10K RPM	150
SAS 2.5" 15K RPM	180
SAS (SED) 15K RPM	180
2.5" SSD	5000+
2.5" FMD	20,000+

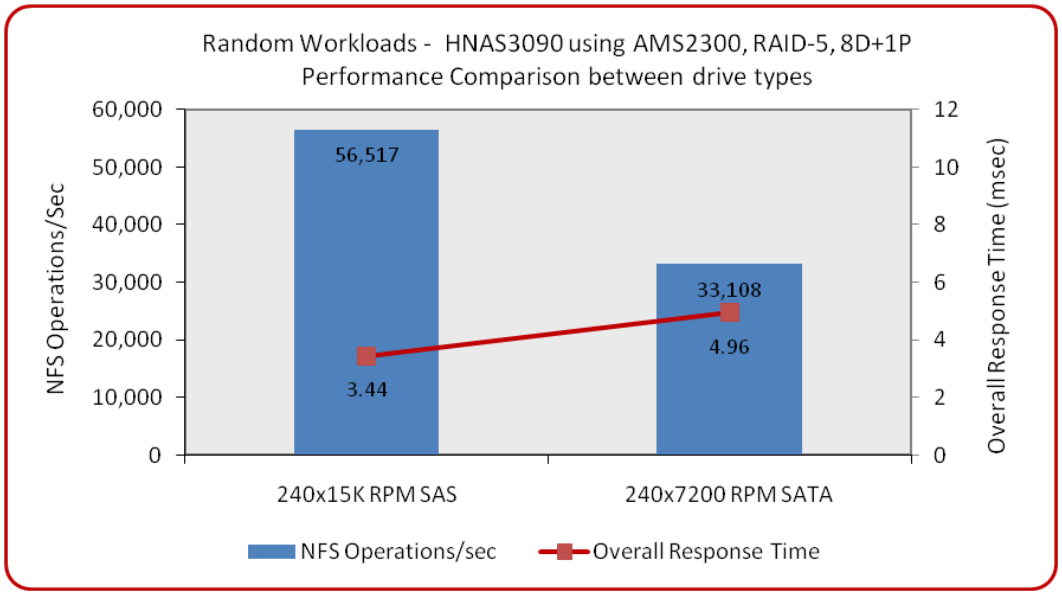
Listed below are some of the observations and recommendations one should be aware of while choosing a disk drive for specific performance requirements:

- When only using the outer 25 percent of the surface in a 15K RPM Disks, the average read seek time can be around 1.8ms and can provide about 250 IOPS. The 100 percent surface average seek time will be about 3.8ms and can provide only about 180 IOPS. *This is also one among the reasons why near full capacity file systems deliver less performance.*
- All write rates are about five percent lower than the read rates due to the higher seek precision requirement.
- In all random read and write cases, distributing a workload across high RPM disks will provide better performance in terms of full random access.

- SATA drives mechanically have much slower seek times and rotation times. As a result, SATA drives can only do about 1/3 as many IOPS per drive as a SAS or Fiber Channel drive. Hence the usage of SATA drive is not recommended when performance requirements are in play.
- *Drives with the same RPM but different capacity deliver about the same performance.* For example, 1TB 7200 RPM SATA drive and 2TB 7200 RPM SATA drive delivers about the same performance. *Within each disk type, a change in capacity does not result in change in performance.*

Figure 1 below illustrates the performance characteristics of SAS and SATA drive types when used in a HNAS 3090, AMS 2300, RAID-5 and 8D+1P environment.

Figure 1: HNAS 3090 using AMS 2300 - Disk Drive performance Characteristics



Solid State Drive (SSD) & Flash Memory Drive (FMD)

A **Solid State Drive (SSD)** is a storage device like a SAS/SATA drive, but uses non-volatile solid-state memory to store the data. A traditional Hard Disk Drive (HDD) is an electro-mechanical device that contains spinning disks and movable read/write heads. But the SSDs use microchips and contain no moving components. The SSD drives have substantially higher throughput advantages for random workloads due to the absence of any mechanical positioning delays. Moreover the SSD drives has the same packaging and interface and it can essentially be substituted anywhere a SAS or SATA HDD could be used.

As the SSD drives have no mechanical positioning, one can expect to see very high random read IOPS capability. But the random write IOPS capability is considerably lower than the read. Also, for large block sequential workloads, there are only fairly small advantages for the SSD drives. Due to their high cost and limited capacity, SSD drives should not be used instead of SAS for predominantly sequential workloads.

Hitachi's custom **Flash Memory Drives** (FMD) could be described as turbo-charged SSDs. Each FMD has much higher performance than an SSD due to much higher internal processing power and degree of parallelism. The first drives are 1.6TB, with 3.2TB drives planned for later release. Inside each FMD is a 4-core processor and ASIC (the Advanced Switch Flash Controller - ASFC) that controls the internal

operations, the four external 6Gbps SAS ports, and the 128 flash memory chips. An SSD drive has one processor and 16 or 32 such flash memory chips. There is a considerable amount of logic in the ASFC that manages the space and how writes are managed. This greatly aids in substantially increasing the write rate for an FMD drive over a standard SSD drive.

Listed below are some of the observations, recommendations and best practices one should consider while using SSD drives on AMS 2000 series systems in an HNAS environment:

- SSD drives and Hard Disk drives can be intermixed in the same tray, but cannot be intermixed within the same RAID Group.
- FMD drives cannot be installed in the regular trays and needs the Flash Memory Unit (FMU) or similar trays. Currently FMD drives are available only for the HUS VM and VSP storage systems (will also be supported in the HUS 100 very soon).
- To take full advantage of all the backend 6Gbps SAS links in the HUS 100 family, one should consider distributing the SSD drives across two standard trays in an HUS 130 and four standard trays in an HUS 150 and create RAID Groups by choosing SSD drives across different trays. Use of two trays (130) enables all 16 6Gbps SAS links, and use of four trays (150) enables use of all 32 6Gbps SAS links.
- When installing all the SSD drives within a single standard tray, the performance capabilities of the drives could be limited due to availability of few backend SAS links and is not a recommended design.
- The SSD and FMD drives have substantially higher IOPS advantages for random workloads due to the absence of any mechanical positioning delays, but a fairly small advantage for the sequential workloads.
- When installed in a high-performance VSP configuration, FMDs have 2-3x greater per drive IOPS horsepower than the SSDs.
- The use of very high performance SSD and FMD drives consumes much more of the internal bandwidth (not to be confused with Sequential throughput) of a storage system than do regular 10K or 15K RPM SAS drives.
- From the storage system perspective, a good rule of thumb to follow is this: each SSD drive uses internal array bandwidth in the ratio of about 12-to-1 to that of SAS drives for 4KB block random read environments. For random writes, this ratio varies considerably by the block sizes, with a 4KB block showing about 9-to-1 ratio of SSD drives to SAS drives.
- Test results suggest that using 30 SSD drives in an AMS 2500 can displace 360 SAS drives for random read environments or 270 SAS drives for random writes in a 4KB block size environment.
- One should not, for example, configure an AMS or HUS system with 30 SSDs and 300 SAS drives with the expectation that both drive types can be driven hard simultaneously.
- In an HNAS environment, to take full advantage of the IO capabilities of these drives, always plan to create many LUNs in each SSD/FMD RAID Group. As there are no moving components within an SSD Drive, creating multiple LUNs in each SSD RAID Group does not impact the performance (nor introduce head thrashing behavior).
- In an HNAS environment, test results suggest that one should create at least 4 LUNs in a RAID-5 (3D+1P) SSD/FMD RAID Group or 8 LUNs in an RAID-5 (7D+1P) SSD/FMD RAID Group.
- As there's no head thrashing issues in a SSD/FMD Drive, after creating multiple LUNs in a SSD/FMD RAID Group, each of those LUNs *should* be assigned to *independent or parallel* HNAS SD Groups (refer to *System Drive Group* section).
- In order to allow enough aggregated host Queue Depth to drive the high performance SSD LUNs, one should distribute the SSD/FMD LUNs across several HUS/HUS VM/VSP and HNAS ports.

Choose the correct RAID Level

The HUS 100 and AMS 2000 family systems support RAID-1, RAID-5, RAID-6 and RAID-1+0. The enterprise family systems (HUS VM, VSP, USP V and older) support RAID-5, RAID-6 and RAID-1+0.

RAID-5 is a group of disks (typically referred to as a RAID Group) with the space equal to one disk used for the rotating parity chunk per RAID stripe (row of chunks across the set of disks). If using an 8D+1P configuration (8 data disks, 1 parity disk), then you get 89 percent capacity utilization for user data blocks from that RAID Group. This is the preferred RAID level for SSDs drives and for SAS disks that see heavy read workloads.

RAID-6 is similar to RAID-5, but with a second parity disk for a second unique parity block. The second parity block includes all of the data chunks plus the first parity chunk for that row. This would be indicated as an 8D+2P construction (80 percent capacity utilization) if using ten disks. This is typically used with SATA disks which have long rebuild times, or where it is critical to never experience a two-drive failure.

RAID-1+0 is both a mirroring and striping mechanism. First, individual pairs of disks are placed into a mirror state. Then, all of these pairs are used in a simple RAID-0 stripe. If using eight disks in the RAID Group, this would be represented as RAID-1+0 (4D+4D) and have 50 percent capacity utilization.

The factors in determining which RAID level to use are cost, reliability and performance. Each RAID type provides its own unique set of benefits. So a clear understanding of your customer's requirements is crucial in this decision.

Another characteristic of RAID is the concept of "write penalty." Each type of RAID has a different back-end physical disk I/O cost, determined by the mechanism of that RAID level. There are additional physical disk reads and writes for every application write due to the use of mirrors or parity. Table 4 below illustrates the tradeoffs between the various RAID levels for write operations.

Table 4: Break down of RAID level Write penalties

RAID Levels	Storage System IOPS per Host Read	Storage System IOPS per Host Write	Storage System IOPS Breakdown per Host Write
RAID-10	1	2	1 data write, 1 mirrored data write
RAID-5	1	4	2 reads (1 data, 1 parity), 2 writes (1 data, 1 parity)
RAID-6	1	6	3 reads (1 data, 2 parity), 3 writes (1 data, 2 parity)

Listed below are the recommendations and best practices one should consider while choosing the RAID level for specific performance requirements:

- The usage of RAID-5 can be recommended for read intensive random workloads and any sequential workloads.
- When using RAID-5 for write heavy workloads, there are four times as many internal controller operations to disk than for RAID-10 and hence not recommended when write performance is the key.
- For write intensive workloads, RAID-10 can often be the least expensive approach, since far fewer disks are required to achieve the same random IOPS levels when compared with RAID-5 or RAID-6.
- The usage of RAID-6 is recommended for the drives that have longer rebuilding times and when using the Hitachi Dynamic Provisioning and Hitachi Dynamic Tiering software.

Choose the optimal RAID Group size

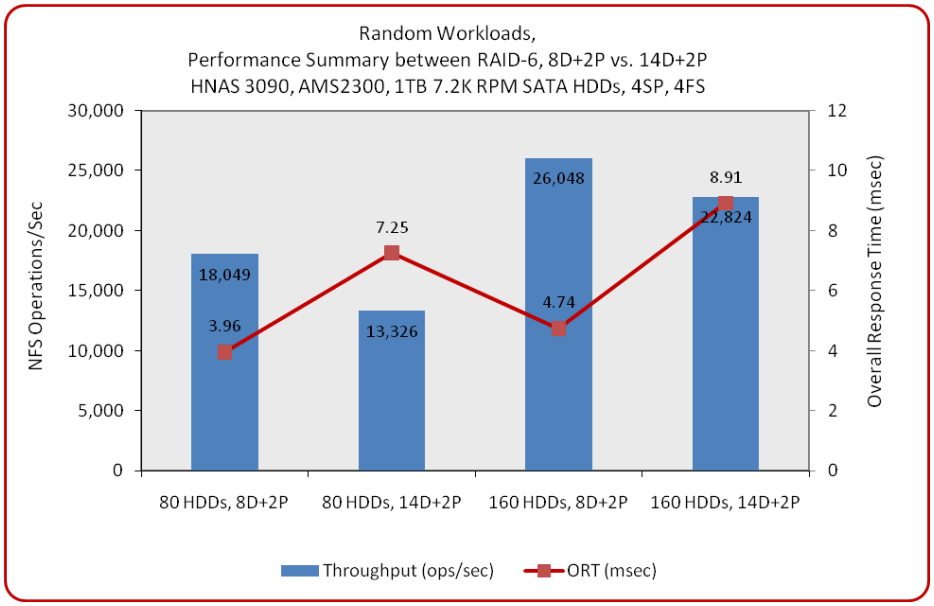
A RAID Group is a logical mechanism that has two basic elements: a virtual block size from each disk (a chunk) and a row of chunks across the group (the RAID stripe). On the HUS 100 and AMS 2000 family, the RAID chunk size defaults to 256KB, and is adjustable to either 64KB or 512KB (per LUN). The stripe size is the sum of the chunk sizes across a RAID Group. This only counts the "data" chunks and not any mirror or parity space. Therefore, on a RAID-6 group created as 8D+2P (ten disks), the stripe size would be 512KB (8 * 64KB chunk) or 2MB (8 * default 256KB chunk). Note that some usage replaces *chunk* with "stripe size," "stripe depth" or "interleave factor," and *stripe size with* "stripe width," "row width" or "row size."

Listed below are the observations, recommendations and best practices one should consider while choosing the RAID Group size:

- The HUS 100 and AMS 2000 family systems attempts to de-stage the data in units of full stripe when possible) (writes are gathered/coalesced at the cache in order to attempt a full stripe write).
- The key challenges with a larger RAID Group size and default chunk size are the partial stripe writes and the related parity overhead. For example, when using RAID-6 14D+2P and 256KB chunk size, the stripe width is 3.5MB. The HNAS I/O size must also be 3.5MB (only possible when using 10.0 release and above) in order to be efficient (full stripe writes, no reading of old data or parity), otherwise partial stripe writes could occur. Hence it is recommended to use smaller chunk size (64KB instead of 256KB from a modular system) and smaller RAID Group size *(example: 8D+2P over 14D+2P).*
- As the default per LUN HNAS Queue Limit is 32, 8D+2P over 14D+2P also results in more number of LUNs and thus results in more aggregated LUN Queue Depth. One could find this beneficial in high performance requirement configurations.
- *One should also understand that the larger RAID Groups takes much more rebuilding/reconstruction time. As a result, the front end host I/O could be impacted during the RAID Group rebuilding time.*

Figure 2 below shows the performance characteristics of RAID-6 (8D+2P) and (14D+2P) in an HNAS environment.

Figure 2: HNAS 3090 using AMS 2300 – RAID-6 (8D+2P) vs. (14D+2P) Characteristics



Number of LUNs per RAID Group

When configuring a storage system, one or more LUNs can be created per RAID Group, but the goal should be to clearly understand what percentage of that group's overall capacity will contain active data. Listed below are the recommendations and best practices one should consider while creating more than one LUN in a RAID Group:

- The maximum supported LUN size (internal LUNs) is 128TB on an HUS 100, 60TB on an AMS 2000 series system and 4TB on a VSP or USP V system. But the HNAS supports only up to 64TB LUN size.
- When several LUNs are carved out of a single RAID Group, their simultaneous use will create maximum seek times on each disk, thus reducing the maximum sustainable small block random IOPS rate to the disk's minimum.
- When a RAID Group has multiple LUNs and when multiple hosts attempt to simultaneously use all the LUNs that share the same physical disks, the seek and rotational latency could become performance limiting factors (also, referred to as disk drive head thrashing).
- When there are multiple LUNs in a RAID Group (that has spinning disks), the HNAS *System Drives* (LUNs) sharing the same physical disks should be placed in the same HNAS *System Drive Group*. (Example: RG-0 has two LUNs 0 and 1. These LUNs should be assigned to HNAS SD Group 0). For more details, refer to the "*System Drive Group"* section.
- It is not recommended to share a RAID Group (that has multiple LUNs) with HNAS and other open hosts.
- In an HNAS environment, it is recommended to create just one LUN per RAID Group when possible.

64KB vs. 256KB LUN Chunk Size (HUS/AMS storage systems only)

A RAID Group is a logical mechanism that has two basic elements: a virtual block size from each disk (a **chunk** - incorrectly referred to as stripe) and a row of chunks across the group (the RAID **stripe**). On the HUS 100 and AMS 2000 family, the LUN chunk size defaults to 256KB, and is adjustable to either 64KB or 512KB on a per LUN basis. The stripe size is the sum of the chunk sizes across a RAID Group. This only counts the "data" chunks and not any mirror or parity space. Therefore, on a RAID-6 group created as 8D+2P (ten disks), the stripe size would be 512KB (8 * 64KB chunk) or 2MB (8 * default 256KB chunk).

One of the challenges with using the default chunk size (256KB) is the I/O block size from the HNAS system. For example, if using RAID-5 8D+1P and the 256KB chunk size, the stripe width is 2MB. The HNAS I/O size (to the storage system) must also be 2MB in order to be efficient (full stripe writes, no reading of old data or parity), otherwise partial stripe writes could occur. Though the largest possible HNAS file system block size is just 32KB and with full ***Superflush*** effects (refer to the *Superflush* section) the largest I/O size could be up to 4MB (when using 10.0 release and above), it is not possible for a HNAS system to write at 4MB I/O size. There are always possibilities of partial stripe writes and the related parity penalties when using the default 256KB chunk size.

But when using RAID-5 8D+1P and the non-default 64KB LUN chunk size, the full stripe width is only 512KB. When better *Superflush* efficiencies are in play and when writes are gathered/coalesced at the storage system cache, the storage system will attempt a full stripe write at most times. *Hence it is recommended to use 64KB LUN Chunk size in a HNAS environment.*

One should also be aware that, in the event of near full capacity or aged or fragmented file systems, the chances for the HNAS to find 4MB (maximum Superflush limit when using 10.0 release and above)) of free contiguous blocks a time are less. During that scenario, Superflush has no effects and 64KB or 256KB LUN Chunk size will make no difference and exhibit similar behaviors.

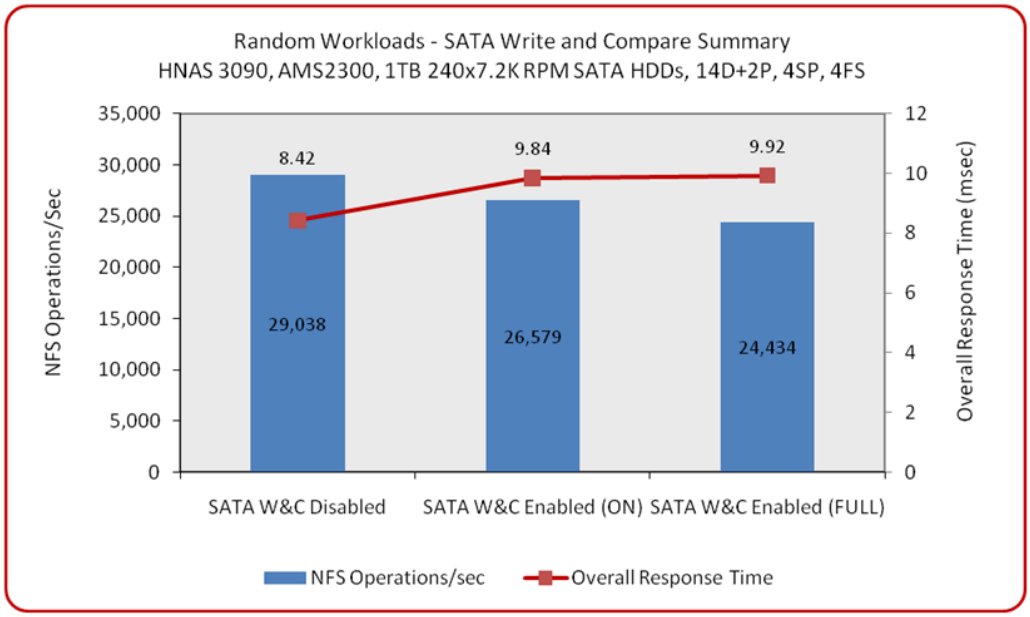
SATA Write and Compare (AMS 2000 family only)

There are three ***Write and Compare*** mode options available when using SATA disks in the AMS 2000 series systems. These include:

- ***Write and Compare Enabled*** (ON) - An optimized read and verification will be performed for write operations to SATA LUNs. This is the default option.
- ***Write and Compare Enabled*** (FULL) - A read and verification will be performed for every write operations to SATA LUNs. A special mode file is required to enable this option (the default mode in AMS1000 systems and below).
- ***Write and Compare Disabled*** (OFF) - The read and verification will not be performed for any write operations to SATA LUNs.

When using 1TB and higher capacity SATA Drives for write intensive workloads, one might consider disabling the SATA Write and Compare mode to achieve slightly better performance levels. Figure 3 shows the performance characteristics of the three SATA Write and Compare modes in an HNAS environment.

Figure 3: SATA Write and Compare Mode Performance Summary



Number of LUNs per Storage System Port

The HUS 100 Family system and the AMS 2000 family system have a ***Port I/O Request Limit*** (**IORL**) of 512 per port. With the introduction of Microcode version 935/A and above, the HUS 100 family now supports up to 1024 per port. This is the limit on the number of aggregate requests on that port for all LUNs visible to the HNAS on that port. The LUN Queue Depth (QD) per port is the number of concurrent threads the HNAS may direct to a single LUN. The default HNAS system drive QD is 32. At 32, an HUS/AMS can accept concurrent I/Os for 16 (=512/32) active LUNs per port before pushing back on HNAS requests.

Every fibre channel path between the HNAS and the storage array has a specific maximum capacity, known as the ***Maximum I/O Request Limit***. This is the limit to the aggregate number of requests being directed against the individual LUNs. For the Virtual Storage Platform system and Hitachi Unified Storage

VM system, this limit is 2048 per port. For the Universal Storage Platform V, this limit is 4096, and is associated with each FED MP. On the 16-port feature, where there are two ports managed by each MP, 4096 is the aggregate limit across those two ports or 2048 per port.

It is very important to ensure that the port IORL is never exceeded. The HNAS system understands the Storage systems port IORL limits and ensures that per port IORL limit is never exceeded. For example, if there are 24 active LUNs mapped to a HUS 100 port and if all the LUNS are busy/active at the same time, the HNAS limits the per LUN max queue depth to 21 (=512/24) on that port. While this ensures that per port IORL is never exceeded, this could also affect performance under certain conditions.

Understanding the performance and capacity requirement should dictate how many active LUNs should be assigned to a single storage port. Environments that have low I/O demands overall could allow a large number of LUNs to be assigned per port without routinely exceeding the queue limit for that port. On the other hand, environments with highly active devices and multiple I/O requests should be designed with fewer LUNs per port, where additional ports will be required in order to satisfy the aggregate I/O demand.

The optimal LUNs per port in a high performance configuration are listed below in Table 5:

Table 5: Port IORL Limits and recommended configuration in an HNAS Environment

Storage System Type	Port I/O Request Limits	Recommended number of active LUNs per port	Maximum LUNs per port (with failover support)
HUS 100	512	16	32
HUS 100 (ver 935/A and above)	1024	32	64
AMS 2000	512	16	32
HUS VM	2048	64	128
VSP	2048	64	128
USP V	2048 (with MP sharing)	64	128

Note: As mentioned above, with the introduction of Microcode version 935/A and above the HUS 100 family now supports up to 1024 per port. But to take full advantage of this enhancement, one has to use HNAS software release 11.2.33xx.xx and above versions.

Number of Fiber Channel Ports

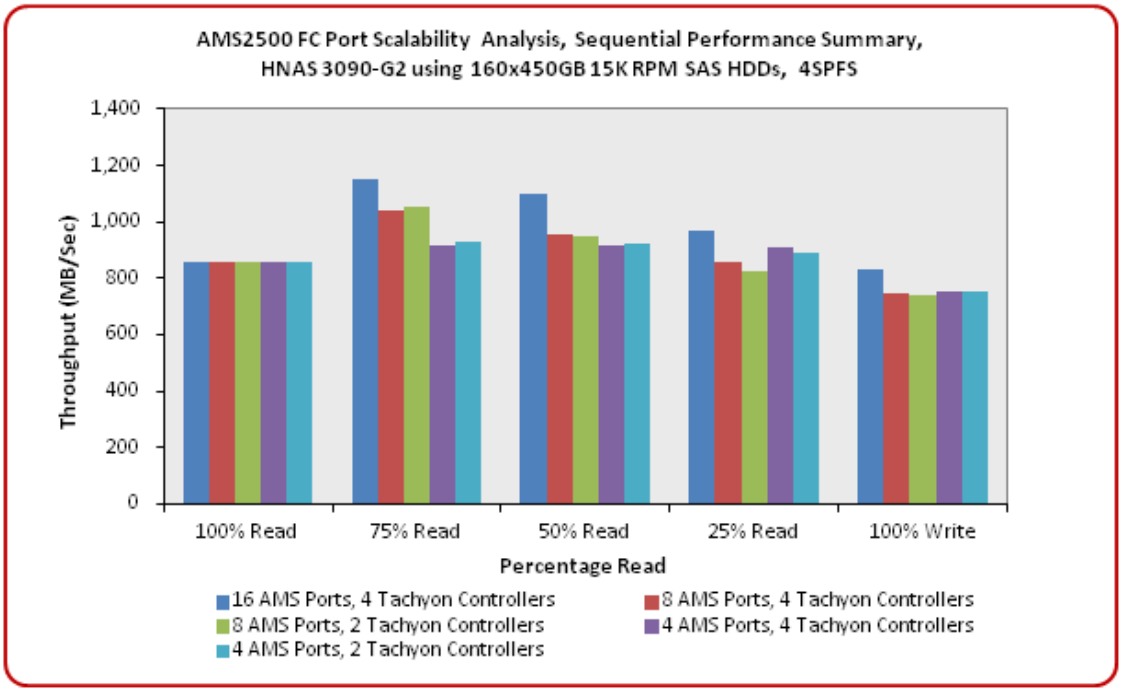
In a switched Fabric (SAN) environment, though not recommended for all environments, it is possible to share a port or ports among several hosts, as is configuring a single host to multiple storage ports. Listed below are some of the observations, recommendations and best practices one should consider while planning for the number of Storage ports in an HNAS environment:

- Host fan-in refers to the consolidation of many host systems into one or just a few storage ports (many-to-one) and has the potential for performance issues by creating a bottleneck at the front-end storage port.
- Having multiple hosts connected to the same storage port only works for environments that have minimal performance requirements.
- While designing a solution, it is important to understand the performance requirements of each host. If each host has either a high IOPS or throughput requirement, it is highly probable that the usage of very few ports will not satisfy the aggregate performance requirements.

- Fan-out allows a host system to take advantage of several storage ports (and possibly additional port Tachyon processors in AMS 2500) from a single or just a few host ports (one-to-many). Fan-out has a potential performance benefit for small block random I/O workloads. This allows multiple storage ports (and their queues) to service a smaller number of host ports.
- Mixing workloads on the ports with large block sequential I/O with small-block random I/O can result in poor performance. In that scenario, the resolution is to use different ports for different data types and I/O profiles.
- The number of LUNs required per HNAS system, number of active LUNs per port and the performance requirements should be considered before finalizing the number of ports.
- For example, if a HNAS system has high performance and/or high capacity requirements, and if that requires 128 LUNs to meet the performance and/or capacity requirements, it will take a minimum of 8 HUS/AMS ports to satisfy that requirement (*512 IORL per port / 32 per LUN QD = 16 active LUNs per port*).
- HNAS Preferred path (refer to sdpath section) should also be configured to balance the load across all the ports.
- Depending on the performance requirements, one should consider using a minimum of 4 or more dedicated storage ports for a HNAS 3090/4000 system and a minimum of 2 dedicated ports for a HNAS 3080 system.
- It is not recommended to share the storage ports with HNAS and other hosts.

Figure 4 below shows the performance characteristics of various numbers of AMS ports in an HNAS environment. Though the chart lists only throughput capabilities for Sequential workloads, one can expect similar behavior for random workloads as well.

Figure 4: Number of Storage Ports analysis in an HNAS environment



Hitachi Dynamic Provisioning

The Hitachi HUS 100 and AMS 2000 Family's **Hitachi Dynamic Provisioning** feature is an internal Open Systems volume management capability. The Hitachi Dynamic Provisioning (**HDP**) feature provides two simultaneous capabilities: ***wide striping*** and ***thin provisioning*****.** The HDP package provides the creation of one or more **Dynamic Provisioning Pools** of physical space per system and for the establishment of **DP Volumes** (DPVOLs, or virtual volumes). Each Pool is created across multiple RAID Groups of the same disk type and RAID level, and then DPVOLs are created against that Pool. These DPVOLs are what the HNAS will see as LUNs or System drives.

The wide striping volume performance feature is the result of the manner in which a Hitachi Dynamic Provisioning Pool is created. A single striped pool of space is established over many RAID Groups, just as is the case with a host-based LVM or HNAS Storage Pools. All of the space per RAID Group is managed as **1GB chunks** of thirty-two contiguous **32MB Pool pages.** These 1GB chunks are allocated as needed to those DPVOLs that are connected to that Pool. The Pool can be viewed as having a chunk size of 1GB across the member RAID Groups. At a given time, the 1GB chunk stays in a single RAID Group. When using the **Advanced Wide Striping Mode**, each of the 32MB Pool pages within a 1GB chunk will come from different member RAID Groups in a round-robin allocation.

The **physical space** for a DPVOL is assigned as needed from the Pool in 1GB chunks of 32 contiguous 32MB Pool pages. These 1GB Pool allocation chunks are evenly distributed (one per RAID Group) across the Pool's RAID Groups by the Pool's allocation mechanism. As a server writes to a DPVOL, and new 32MB Pool pages are required, these are pulled from the DPVOLs local working set of Pool pages from the most recently allocated 1GB chunk. When a new chunk is needed to supply the next group of 32 32MB pages, an additional chunk will be round-robin allocated to that DPVOL from a different RAID Group in that Pool.

In the VSP or USP V enterprise storage systems, the DPVOL is allocated as 42MB pages (physical space) from a Pool as the HNAS writes to blocks in that volume. The physical space assigned as needed from the HDP Pool to a V-VOL Group container is evenly distributed across the Pool Volumes by the mechanism of the Pool's Free Page List. As the HNAS writes to a DPVOL, and new pages are required, these are assigned to the V-VOL Group for use by that DP-VOL.

Whereas in the HNAS, each storage pool is typically created by striping four or more system drives (LUNs). When creating a Storage Pool, the HNAS stripes the system drives together using a 4MB stripe. When the I/O is issued to the Storage Pool, it is in turn sent to the underlying Stripe. Depending on the active data set size, all the system drives within that stripe are used to achieve the maximum performance. Storage Pools can be expanded as additional SDs are created in the storage subsystem, and grow to a maximum capacity of 256 TB (or, 1PB when using 10.1 release and above), but the maximum number of system drives within a single stripe set is 32. When additional system drives are added to the existing Storage Pool, the HNAS creates another stripe set using a 4MB stripe.

When comparing the HDP Wide Striping feature with the HNAS Storage Pool striping feature, due to the default HNAS stripe size of just 4MB, *the HNAS handles the stripe and the underlying LUNs much more efficiently than the HDP*. In an HNAS and AMS 2000 family environment, one could notice the performance benefits when using the regular LUNs (non-HDP) and the default HNAS Storage Pool striping feature over DPVOLs and HDP Wide Striping feature.

However there is no such behavior when using DPVOLs from an HUS 100 family storage system or the HUS VM system. Test results indicate the performance levels between HDP and Non-HDP are similar. One could notice the performance characteristics of using DPVOLs from HDP Wide Striping feature and HNAS default Storage Pool Striping (HNAS Striping on top of HDP Striping) when using HUS 100 family storage system from the below Figure 5.

Figure 5: HUS 100 family HDP vs. Non-HDP performance summary in an HNAS environment

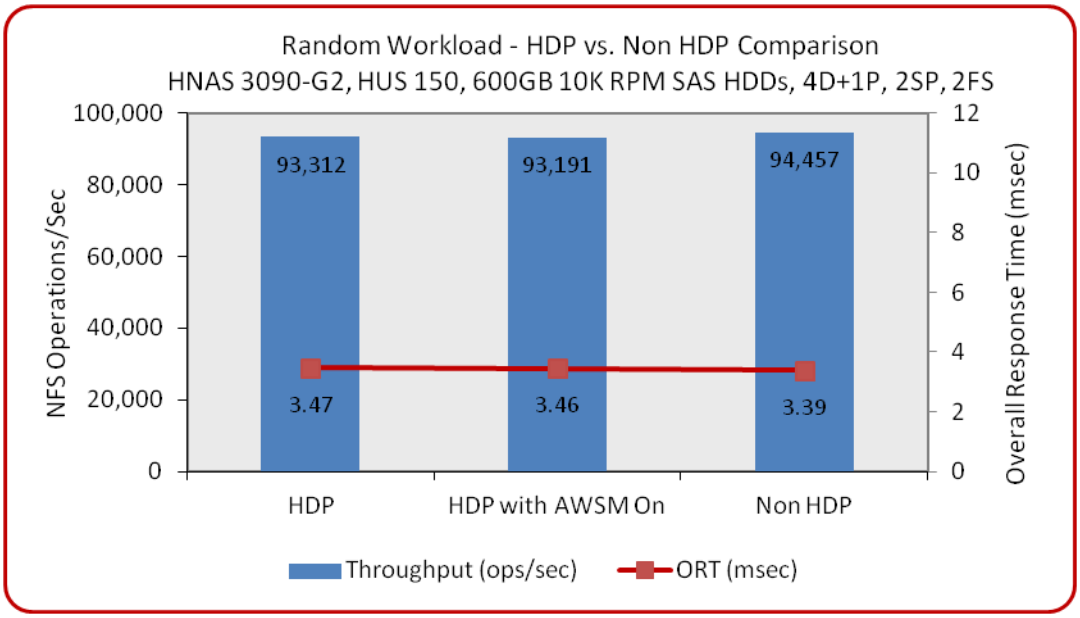
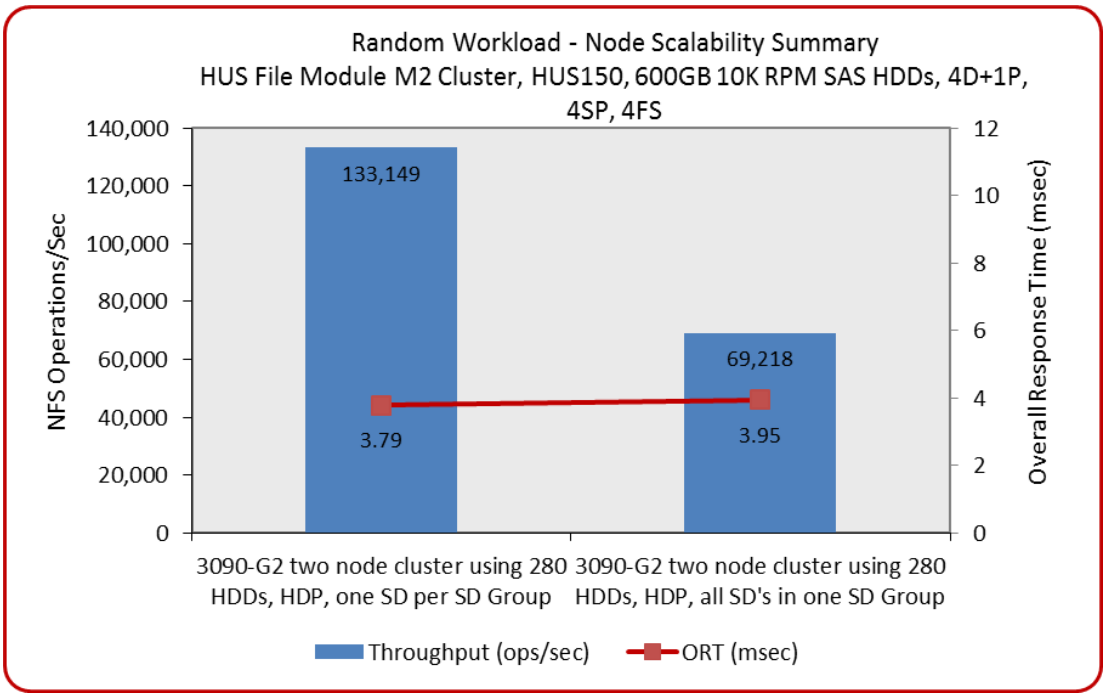


Figure 6: System Drive Group performance summary



Listed below are some of the observations, recommendations and best practices one should consider while using HDP DPVOLs in an HNAS environment:

- While creating a HDP Pool, use the same disk type and RAID Level for every Pool Volume within that Pool. It is recommended that there be 32 or more disks per HDP Pool in order to have an adequate amount of disk IOPS available to the HDP Pool.
- When using non-HDP volumes, LUNs hosted on the same physical disks should be placed in a single HNAS System Drive Group (SD Group). When using DPVOLs from the Hitachi Dynamic Provisioning feature, from the HNAS or administrator perspective it is not possible to find out at which

RAID Group does the 1GB chunk is currently physically located at. Hence when using HDP DPVOLs in a HNAS system, one should place each system drive (DPVOLs) in ***different or independent*** SD Groups. Refer to figure 6 above to understand the impact of incorrect SD group configuration.

- When using AMS 2000 family system in a HNAS environment, in most of the test cases the standard LUNs (Non-HDP) and the default HNAS Storage Pool striping mechanism outperforms use of DPVOLs and HDP Wide Striping by about 10%. The multiple striping layers (HNAS striping on top of HDP striping) affects the overall performance levels from the HNAS perspective. *However there is no such behavior when using DPVOLs from an HUS 100 family storage system and results indicate the performance levels between HDP and Non-HDP are similar.*
- Never create a storage pool using just one DPVOL. For file system protection mechanisms, HNAS stores critical file system metadata information on multiple system drives within a storage pool. In the event of a failure or corruption, this critical data can be copied from one drive to another. But when only one system drive is used within a storage pool, there are no protection mechanisms available and conceivably puts data at risk.
- Always monitor and ensure that the HDP Pool is not running out of free space. One way to safe guard a HDP pool dedicated to the HNAS is to not allow over-provisioning (over-subscribe). When a HDP Pool runs out of free physical space, the HNAS will experience write errors and will bring down the associated HNAS file systems and potential data loss might occur.
- In an HUS 100 and AMS 2000 series system, it is mandatory to set the *Over-provisioning Limit Alert* threshold to 100% (or less) (specific to HNAS environment).
- In an HUS 100 and AMS 2000 series system, set Advanced Wide Striping Mode (AWSM) on. AWSM makes sure that the 32MB pool pages within each chunk are distributed across the different RAID Groups.
- Set SOM 917 on (when available). SOM 917 makes sure that the poo/tier is rebalanced across the Parity Groups rather than the pool volumes (LDEVs).

Hitachi Dynamic Tiering

Hitachi Dynamic Tiering (HDT) is a new feature of the Hitachi Virtual Storage Platform (VSP) and is available in the Hitachi Unified Storage VM (HUS VM) and Hitachi Unified Storage 100 family (HUS 100) as well. HDT builds on the strengths of Hitachi Dynamic Provisioning (HDP), which was first introduced on Hitachi's USP-V platform. HDT provides the following features:

- **Thin Provisioning**, which eliminates wasted storage capacity by incrementally allocating physical disk space as the host stores data;
- **Wide Striping** across a pool of disks, which eliminates disk bottlenecks often associated with traditional storage management;
- **Automated Tiering**, which improves performance by storing frequently accessed data on fast disks, and reduces cost by storing rarely referenced data on inexpensive, large-capacity disks.

HDT seeks to improve performance and reduce cost by optimizing disk usage. To accomplish this, HDT monitors back end disk access to 42MB pages for a period of time called a monitoring cycle. The HDT monitoring method may be simple (**Period** mode) or may employ a weighted moving average of back end I/Os per hour (**Continuous** mode). At the end of each monitoring cycle, HDT analyzes its monitoring data and creates a page access frequency distribution. Based on the frequency distribution, HDT establishes Tier ranges that are used to guide each page to the most appropriate Tier in a process called relocation. Frequently referenced pages are moved to Tier 1 (top tier) to improve performance. Less frequently referenced pages are relocated to higher-capacity, more economical media (low tier) to reduce cost. Relocation normally runs concurrently with HDT monitoring. Through continual monitoring and relocation, HDT attempts to keep as many pages as possible in upper Tiers to improve performance, but also tries to keep each Tier at or below 60% of its performance capacity, to improve response time and allow for

overhead operations. HDT requires multiple monitoring cycles to achieve and maintain the most efficient distribution of data across Tiers.

An HDT Tier is composed of Pool Volumes having the same underlying disk type and RAID type**.** HDT supports a maximum of three Tiers per HDT Pool. The highest-performing media installed in a Pool is referred to as Tier 1. For example, Tier 1 of an HDT Pool might be composed of LDEVs allocated on two RAID-6 6D+2P SSD/FMD Parity Groups. A Tier 1 thus configured would deliver the IOPS power of 16 SSDs. Lower Tiers would be composed of disks with higher storage capacity but lower IOPS power. Tier 2 would typically contain 10K SAS disks and is often considered a good place for new page allocation, since it has a good balance between performance and cost. Newly allocated pages can be monitored while on Tier 2, until HDT has enough information to decide whether movement to a higher or lower Tier is warranted. Tier 3 would be composed of disks with the most storage capacity but the least IOPS power. Typical examples of Tier 3 disks would include SATA, 7KSAS, and/or virtualized external storage. Pages with very low levels of activity will ultimately be moved to Tier 3, where they will remain accessible at the lowest possible cost.

A DP-VOL is a dynamically provisioned logical device (volume or LUN) that is presented to HNAS as system drives. DP-VOLs are sometimes called VVOLs or just LDEVs in VSP/HUS VM terminology. To the HNAS, HDP and HDT DP-VOLs are indistinguishable from each other. However internally, the page management for HDT DP-VOLs is more complex than for HDP. An HDT DP-VOL may have its pages distributed across up to three Tiers of disk. A new page for an HDT DP-VOL will be acquired from the highest Tier with available space.

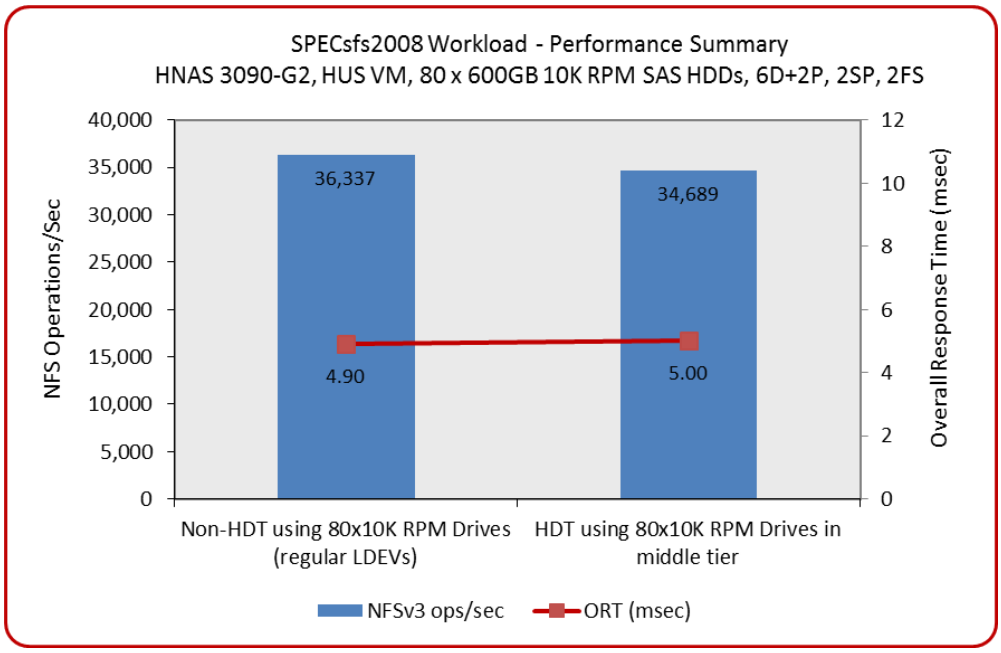
In order to optimize page placement on the Tiers, HDT must monitor data access patterns for a period of time known as a **monitoring cycle.** It is important to understand that HDT does not attempt to optimize page placement in real-time. Rather, HDT watches historical access patterns to decide which pages to move to which Tier. HDT is primarily concerned with improving performance and reducing cost by optimizing disk usage. HDT does not concern itself with the number or type of host operations on the FED ports. Rather, HDT is only concerned with host operations that are ultimately translated into physical disk accesses to each page. Essentially, HDT counts the number of times that the disks underlying each page have been read or written during the monitoring cycle. At the end of each monitoring cycle, this measurement of disk I/O is standardized into I/Os per hour, or IOPH for each page. HDT counts only certain types of disk operations. Only read data and write data disk operations are included in IOPH measurements.

After a monitoring cycle is complete and Tier ranges have been established, HDT's next task is to **relocate** pages to the most appropriate Tier. Relocation is expected to consume less than 5% of the VSP's CPU cycles. Host (in this case, HNAS) I/O is given priority over relocation processing. Lab testing to date has shown no detectable difference in host (HNAS) I/O performance to a Pool with relocation active, when compared to the same workload on the same Pool with no relocation occurring. The pace of relocation varies with the RAID type and disk type of the source and target Tiers. The target Tier tends to be the limiting factor, since RAID writes require more physical disk operations than reads. For example, reading from a RAID-5 Tier and writing to a RAID-6 Tier is likely to be slower than the reverse. Contention for the same resources between host I/O and relocation also affects relocation performance.

The general guideline is that relocation can be expected to move about 3 TB per day across all Pools, but actual observations have varied widely. Relocation has been measured as slow as 4 MB/second (345 GB per day), when the pages being relocated are also extremely busy with host I/O. On the other hand, when there is minimal competing host I/O load, relocation as fast as 85 MB/second (7 TB per day) has been observed.

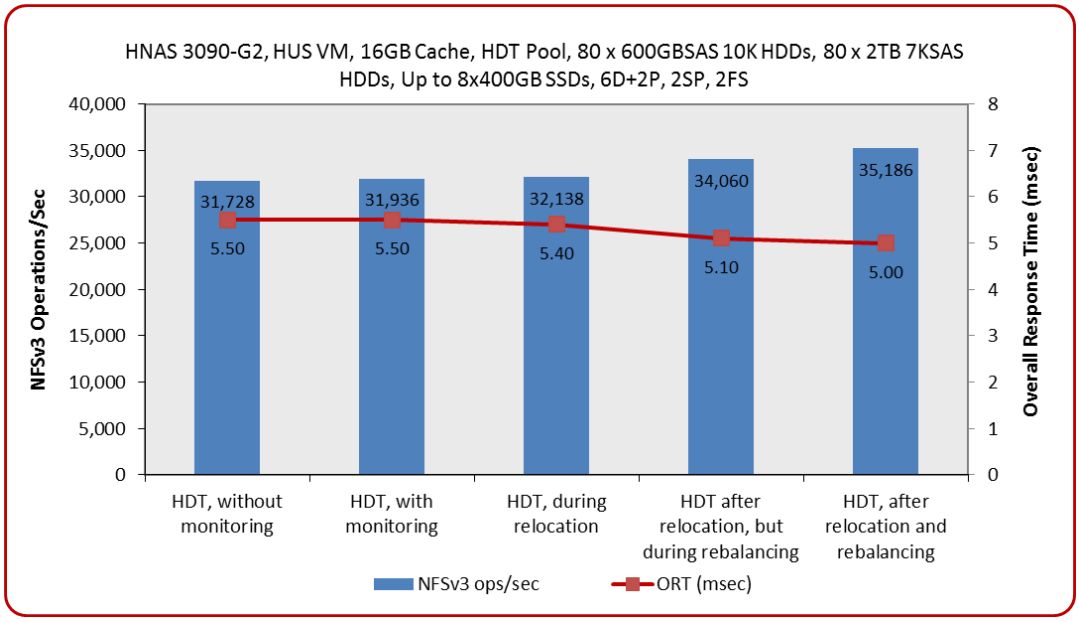
One could notice the performance characteristics of using DPVOLs from HDT and HNAS default Storage Pool Striping when using HUS VM storage system from the below Figure 7. In the below example, during the HDT test there was no HDT monitoring, relocation or rebalancing process running at the background (characteristics similar to an HDP pool).

Figure 7: HUS VM HDT vs. Non-HDT performance summary in an HNAS environment



However, depending on the workload characteristics one could see the performance benefits of HDT after the active page relocations. The below test was conducted to measure the performance characteristics of metadata intensive random workloads when using HDT and without page monitoring, with monitoring, during page relocation, during tier rebalancing (if any) and after page relocation.

Figure 8: Random Workloads performance Summary - During various HDT operations



Listed below are some of the observations, recommendations and best practices one should consider while using HDP DPVOLs in an HNAS environment:

- The use of HDP and HDT does not make the disk drives perform any faster. HDP/HDT is about creating wide stripes across many RAID Groups (just as with a HNAS storage pool) to avoid the effects of workload skew when using many individual RAID Groups.
- It is recommended that the new HDT Pools be configured in RAID-6.
- It is recommended that there be minimum 32 or more disks per HDT Tier (SSDs are an exemption) to achieve the performance target (an adequate amount of disk IOPS in the HDT Pool). However, the number of disks in a Pool/Tier must be determined by estimating the sustained peak IOPS required along with the RAID level write penalty factor.
- Set SOM 917 on (when available). SOM 917 makes sure that the poo/tier is rebalanced across the Parity Groups rather than the pool volumes (LDEVs).
- Set SOM 896 on. SOM 896 enables background page formatting. With 896 off, pages are formatted at the time they are allocated, which incurs a performance penalty.
- Set SOM 901 on. SOM 901 makes sure that the pages are not demoted when the performance utilization on an SSD tier exceeds 60%. As the response times from the SSD are relatively good, the performance utilization on the SSD tier can exceed the 60% recommended limit.
- If an HNAS system experience storage response time degradation in the environments with frequent relocation and rebalancing, consider turning on SOM 904. SOM 904 reduces the pace of relocation/rebalancing to one page per second (42MB/s).
- One may find it beneficial to reserve more than the default 8% of capacity in Tier 2 for new page allocation, to encourage more new pages to be created in Tier 2.
- Note that the HDP and HDT features are designed to even out skewed host loads across the member RAID Groups within a HDP/HDT Pool, where it outperforms the non-HDP and non-HDT environment. When using uniform workloads (a 100% constant load against every LUN using all host paths evenly) in a controlled test environment, the usage of HDP or HDT are typically expected to deliver slightly lower performance due to its overhead.
- An HDT Pool hosting HNAS System Drives (SDs) should never be over-provisioned. In order to ensure that the HDT Pool is never over-provisioned, the subscription limit for each HDT Pool shall be set to 100% or less.
- The size of the DP-VOLs must never change. The HNAS file system cannot recognize changed DP-VOL capacities.
- Except for the HNAS Tiered File Systems, all the DP-VOLs used within a HNAS Storage Pool should have the same performance capabilities.
- It is recommended to use a minimum of 2 system drives in a HNAS storage pool. For file system protection mechanisms, HNAS stores critical file system metadata information on multiple system drives within a HNAS storage pool. When only one system drive is used within a storage pool, there are no protection mechanisms available and conceivably puts data at risk.
- For both random and sequential workload tests (uniform, not skewed), the results showed comparable performance between non-HDT and HDT (a tiered pool without any monitoring or relocating operations). Though the Non-HDT configuration marginally outperformed HDT configuration by up to 7%, overtime the performance results of HDT after the tier relocation are higher.
- For metadata intensive random workloads, the results showed comparable performance between non-HDT and HDT (a tiered pool without any monitoring or relocating operations). . Though the Non-HDT configuration marginally outperformed HDT configuration by up to 5%, overtime the performance results of HDT after the tier relocation are much higher.
- For all the read intensive workloads, the overall performance from the HDT pool improved overtime after the relocation and rebalancing process was complete.

- During one of the metadata intensive random workload tests, the HDT has promoted hot pages (~358GB) to the SSD tier and demoted the cold pages (~4.98TB) to the 7KSAS tier. In spite of demoting ~4.98TB of cold pages to the 7KSAS tier, we were still able to achieve higher ops/sec (by 10%) and better response times (by 9%).
- When using HNAS Tiered File Systems with various HDT, non-HDT and hybrid (HDT and non-HDT) configurations, there were no noticeable differences in terms of ops/sec and response times. The performance differences between any of the configurations were lower than 5%for both ops/sec and response times.

HNAS Concepts & Best Practices

Like the backend storage system design and configuration, it is equally important to know how the HNAS system will address the customer's specific business and performance needs. The models of HNAS configured, including the number of FC connections, tunable parameters, and protocol specific considerations etc are all highly important to the whole solution.

Choose the correct HNAS System

First released in July 2013, the HNAS 4000 family includes the models 4060, 4080 and 4100. Models 4060 and 4080 are the replacements for the previous Hitachi NAS Platform models 3080 and 3090 respectively. Model 4100 is the replacement for the previous Hitachi High Performance NAS Platform model 3200. The HNAS 4000 systems have much higher performance than the previous generation and incorporate several significant hardware changes.

The 4000 series delivers best-in-class performance, scalability, clustering with automated failover, 99.999% availability, non-disruptive upgrades, smart primary deduplication, intelligent file tiering, automated migration, 256TB file system pools, a single namespace up to the maximum usable capacity and are integrated with the Hitachi Command suite of management and data protection software. The 4000 series nodes can scale up 16PB of usable data storage; support 10GbE LAN access and 8Gbps FC storage connectivity.

The previous generation Hitachi High-Performance NAS Platform, model 3200, offer best-in-class performance, scalability, clustering up to eight nodes, up to 125 or 128 file systems, maximum 16PB usable capacity, two 10GbE Ethernet ports or six 1GbE Ethernet ports, eight 4Gbps Fibre Channel ports etc.

The previous generation Hitachi NAS Platform, models 3090-G2 or 3090-G1, offer best-in-class performance, scalability, clustering up to four nodes, up to 125 or 128 file systems, maximum 8PB usable capacity, two 10GbE Ethernet ports, six 1GbE Ethernet ports, four 4Gbps Fibre Channel ports etc.

The previous generation Hitachi NAS Platform, models 3080-G2 or 3080-G1, supports clustering up to two nodes, up to 125 or 128 file systems, maximum 4PB usable capacity, two 10GbE Ethernet ports, six 1GbE Ethernet ports, four 4Gbps Fibre Channel ports etc. The key performance differentiators (between the models) are listed below in Table 6:

Table 6: Performance characteristics of HNAS systems

HNAS Model	Random performance (SPECsfs) - NFSv3	Throughput in MB/s
4100	145,000	Up to 2,000
4080	105,000	Up to 1,500
4060	74,000	Up to 1,000
3200	195,000*	Up to 1,400
3090 with Performance Accelerator Feature	95,000	Up to 1,450
3090	74,000	Up to 1,150
3080	41,000	Up to 675

* indicates SPECsfs v97_R1

Tachyon Processor Overview and Limitations

The Tachyon QE processor in the HNAS 4000 provides the interface between the DI FPGA and the backend storage system. The Tachyon processor provides full duplex, non-blocking bandwidth to each of the four 8Gbps Fibre Channel ports. The per port read and write bandwidth of the FC ports are listed in the table below. Although the 4060, 4080 and 4100 use the same number of FC ports and the same potential throughput bandwidth, the actual performance of each model varies based on other internal factors.

The Disk Interface in the HNAS 3090/3080 systems and the Storage Interface Module in the HNAS 3200 system is used for connectivity to the storage system on the back-end Fibre Channel SAN. At the core of each disk or storage interface is the Tachyon processor providing full duplex and dedicated bandwidth to each Fibre Channel port.

The HNAS 3200 uses two Tachyon QX4 Chipsets in the SIM-3 interface. The Host Ports (or hports) 1 to 4 share the first Tachyon QX4 Chipset and the hports 5 to 8 share the other Tachyon controller. These systems have a per port bandwidth limitation of 180 MB/s for writes and 400 MB/s for reads. One should also be aware that the 3200 systems have a maximum throughput limit of about 1.4 GB/s (for mixed workloads).

The early shipments of HNAS 3090/3080 G1 systems use a Tachyon QX4 Chipset. All the four hports share the same Tachyon QX4 Chipset. These systems have a per port bandwidth limitation of 180 MB/s for writes and 400 MB/s for reads. One should also be aware that the 3090-G1 systems have a maximum throughput limit of about 1 GB/s (for mixed workloads).

The HNAS 3090/3080-G2 systems use a Tachyon QE4+ Chipset. All the four hports share the same Tachyon QE4+ Chipset. These systems have a per port bandwidth limitation of 360 MB/s for writes and 400 MB/s for reads. One should also be aware that the 3090-G2 system has a maximum throughput limit of about 1.1 GB/s (for mixed workloads). Listed below in Table 7 are the Tachyon processor bandwidth limitations break down per HNAS model (without the Performance Accelerator feature):

Table 7: Tachyon throughput limits break down by various HNAS models

HNAS Model	Tachyon Chipset Details	Per port, Read Limits	Per Port, Write Limits	System Peak Limits
4060	QE8	800 MB/s	600 MB/s	1 GB/s
4080	QE8	800 MB/s	600 MB/s	1.5GB/s
4100	QE8	800 MB/s	600 MB/s	2 GB/s
3200	Two QX4	400 MB/s	180 MB/s	1.4 GB/s
3100	QX4	400 MB/s	180 MB/s	700 MB/s
3090-G2	QE4+	400 MB/s	360 MB/s	1.1 GB/s
3090-G1	QX4	400 MB/s	180 MB/s	1 GB/s
3080-G2	QE4+	400 MB/s	360 MB/s	700 MB/s
3080-G1	QX4	400 MB/s	180 MB/s	700 MB/s

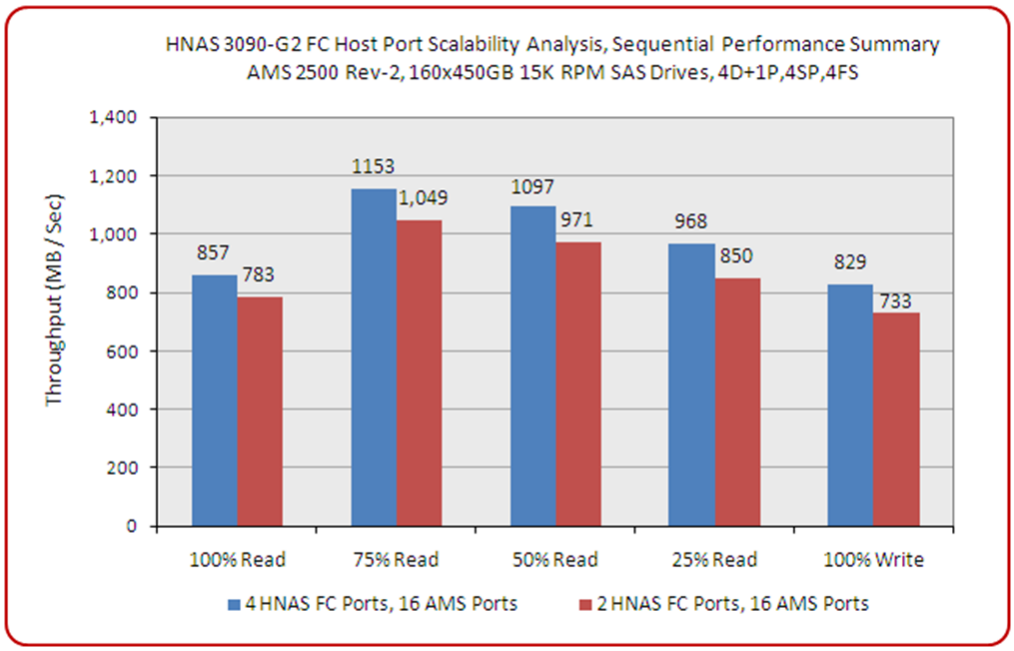
Number of HNAS Fiber Channel Ports

Listed below are some of the observations, recommendations and best practices one should consider while planning for the number of Host ports in an HNAS system:

- While designing a solution, it is important to understand the performance requirements of each node. If each node has either a high IOPS or throughput requirement, it is highly probable that the usage of very few Host Ports (or, “hports”) will not satisfy the aggregate performance requirements.
- In a high performance environment, one should take advantage of several hports (and thus additional Tachyon chips in case of 3200). The usage of several hports has performance benefits for sequential workloads as well.
- Most general purpose systems and random workloads can be satisfied by using a minimum of 2 hports on 4000/3090/3080 systems and 4 hports across the two Tachyons on the 3200 systems.
- In a high performance Sequential workload environment and considering the throughput limits of different Tachyon controllers (refer to Table 7), one should use all the 4 hports on 3090/4080/4100 systems, a minimum of 4 hports across the two Tachyons on the 3200 systems and 2 hports on the 3080/4060 systems.

Figure 9 below shows the performance characteristics of either 2 or 4 hports in a HNAS 3090 system.

Figure 9: Number of Host Ports in a HNAS system - Performance Analysis



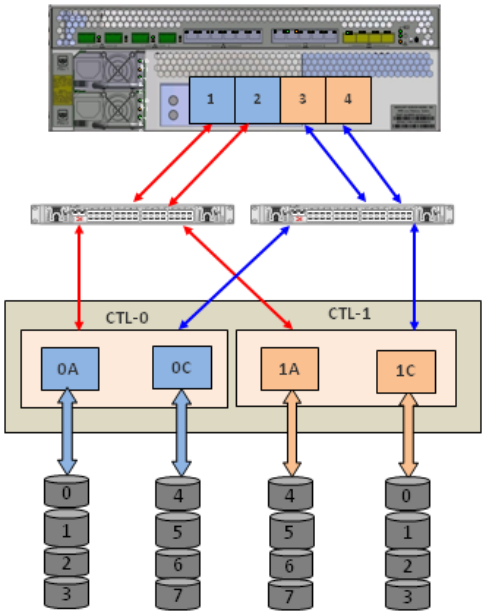
Sdpath

The HNAS systems are equipped with a multi-path driver and can use alternate paths to LUNs from the backend storage systems. This means that if a HNAS system is connected to a pair of controllers (example, AMS 2500) that have two LUNs presented, the system will attempt to choose a port from AMS CTL-0 as a primary path to the first LUN and then the other port from CTL-1 as a primary path to the second LUN. The system will also dynamically change the paths in the event of a path failure. The *sdpath* command is used to view or change the preferred storage path by the node or the cluster. When possible, any changes to the preferred path would take effect immediately on all the nodes in a cluster.

While the automatic selection algorithm is often good enough for most systems, the manual configuration will ensure better load distribution and could lead into better performance in few complex configurations. *For better load distribution and optimal performance, it is either recommended to manually setup preferred host ports and target ports to all the system drives in a node/cluster or check and make sure the load is auto balanced across the host ports and target ports automatically.*

Figure 10 below shows a simple configuration diagram using a single HNAS 3090 system and an AMS 2000 storage system.

Figure 10: Configuration diagram of a simple HNAS solution



One can logically notice from the above figure that hports 1, 2 are zoned to AMS ports 0A, 1A. Similarly hports 3, 4 are zoned to AMS ports 0C, 1C. For the above example, the below or similar preferred path concept is recommended:

- sdpath -h 1 -t 0A 0
- sdpath -h 1 -t 0A 1
- sdpath -h 3 -t 1C 2
- sdpath -h 3 -t 1C 3
- sdpath -h 2 -t 1A 4
- sdpath -h 2 -t 1A 5
- sdpath -h 4 -t 0C 6
- sdpath -h 4 -t 0C 7

Sdpath (when using HUS 100 family only)

The HNAS systems are equipped with a multi-path driver and can use alternate paths to LUNs from the HUS storage systems. This means that if a HNAS system is connected to a pair of controllers (example, HUS 150) that have two LUNs presented, the system will attempt to choose a port from HUS 150 CTL-0 as a primary path for the first LUN and then the other port from CTL-1 as the primary path for the second LUN.

The HUS 100 family storage system also provides preferred path information via a SCSI VPD (Virtual Product Data) page E0. The HNAS uses the information in the page E0 output and implements the preferred path accordingly. The HNAS polls for the E0 information every 5 minutes to determine if the ownership of a LUN has changed either due to HUS 100 controller load balancing mechanism or path failure. In the event of any storage side ownership changes, the HNAS system will also dynamically change the paths and establish the preferred path to a LUN.

However beginning the HUS 100 microcode version 09.35A and HNAS version 11.2.32xx, due to new enhancements the E0 procedure has been supplanted and replaced by a far robust mechanism (SCSI Unit Attention from the storage array). To take full advantage of this enhancement, one has to set the "HNAS Option Mode" in the HUS host groups that are used by the HNAS system.

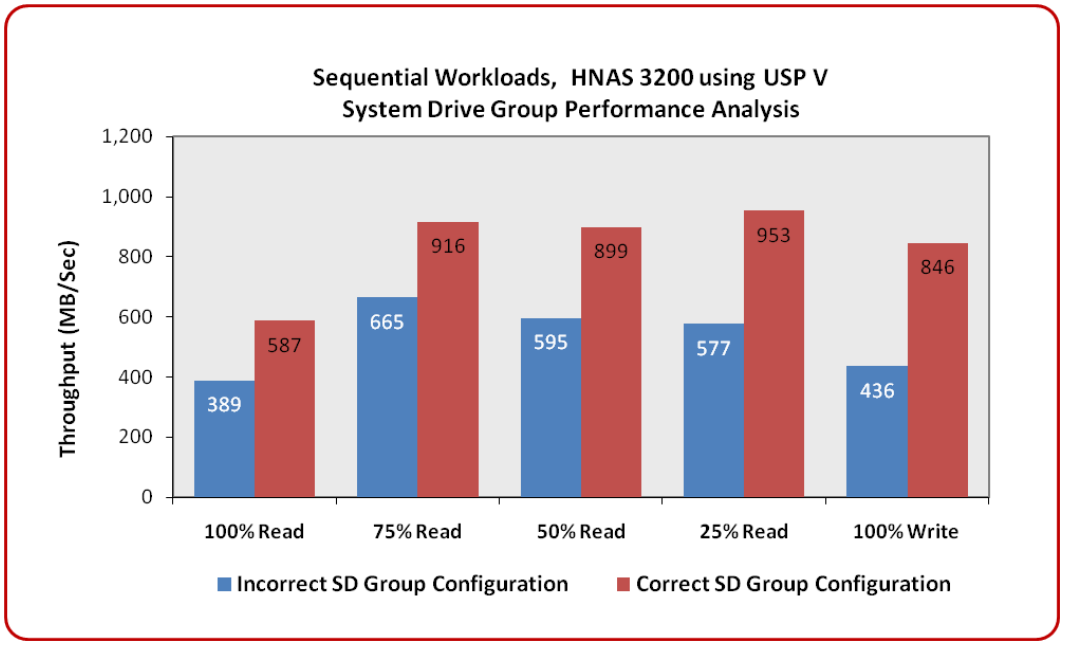
For these reasons, when using HUS 100 family storage system one need ***not*** setup sdpath manually and use the automatic selection algorithm. The *sdpath* command can be used to view or rebalance (if needed) the preferred storage path by the node or the cluster. Any changes to the preferred path would take effect immediately on the other node in the cluster.

System Drive Groups

The ***System Drive Group*** (**SDG**) is a mechanism for telling the HNAS which of the System drives are sharing the same Disks (spindles) within a RAID group in the storage subsystem. Thus it can avoid writing to several members of the same RAID group at once and thus decreasing the disk drive head movement. Let's say for example a RAID Group has 2 LUNs. Normally the first half of each physical disk is dedicated to the first LUN and the second half of each disk is dedicated to the second LUN. If the HNAS writes to both the LUNs at the same time, this would result in continual head movement between the first and second halves of each physical disk and thus would affect the performance.

To avoid this situation from occurring, if a spinning disk RAID Group has 2 LUNs, then one should place both the LUNs in the same HNAS system drive group, thus telling HNAS that both the LUNs are hosted on the physical disks and that it shouldn't write to the both the LUNs at once. Now the SDG maintains only one write cursor per file system hosted on those system drives. When a spinning disk RAID Group has two LUNs and if you place them in different system drive groups by mistake, this would falsely tell the HNAS that the LUNs don't share the same physical disks and would affect the performance. Similarly, when there are multiple RAID Groups (each RG on unique physical disks) and each RAID Group has one LUN and if you place all the LUNs in a single drive group by mistake, this would also falsely tell the HNAS that the LUNs are sharing the same physical disks and would affect the performance. Figure 11 below shows how an incorrect SD Group configuration can affect the performance:

Figure 11: SD Group Analysis



Listed below are some of the observations, recommendations and best practices one should consider while configuring the SD Group in a HNAS system:

- LUNs hosted on the same physical disks within a spinning disk RAID Group should be placed in a single HNAS system drive group. For example, if there are 4 LUNs in a RAID Group, then all the 4 HNAS SDs should be placed in a single SD Group.
- When using DPVOLs from the Hitachi Dynamic Provisioning or the Hitachi Dynamic Tiering features, one should place each system drives (DPVOLs) in ***different or independent*** SD Groups.
- When using LUNs from SSD/FMD Drives in a HNAS system, as there are no moving components and thus no related head thrashing issues in these drives, one should create multiple LUNs per SSD/FMD RAID Group. Refer to the SSD/FMD section for more details.
- As there are no head thrashing issues in SSD/FMD drives and RAID groups, the multiple LUNs in a SSD RAID Group *should* be assigned to ***different/parallel*** SD Groups to achieve peak performance levels. An incorrectly configured SD Group would affect the overall performance levels of an HNAS system.

Storage Pool

Each ***Logical Unit*** (**LUN**) presented from a storage array to the HNAS is called a ***System Drive*** (**SD**). A ***Storage Pool*** consists of one or more system drives and is the logical container for one or more file systems. When creating a Storage Pool, the HNAS creates a RAID-0 stripe across the system drives using a 4MB stripe size. When the I/O is issued to the Storage Pool, it is in turn sent to the underlying Stripe. Depending on how big the current dataset is, all the system drives within the stripe are used to in order to achieve the maximum performance. The maximum number of system drives within a single stripe is 32. Storage Pools can be expanded as additional SDs are created in the storage subsystem, and grow to a maximum capacity of 1PB. When additional system drives are added to the existing Storage Pool, the HNAS creates another stripe set using a 4MB stripe.

A Storage Pool can hold up to 125 (in cluster systems) or 128 (in single node systems) file systems, centralizing and simplifying management of its component file systems. Storage Pools are made up of multiple small allocations of storage called “chunks.” The size of the chunks in a Storage Pool is defined when the Storage Pool is created. A Storage Pool can contain up to a maximum of 16,384 chunks. In turn, an individual file system can contain up to a maximum of 1023 chunks. *However with the introduction of 11.2 release, a storage pool and file system can now each contain up to 60,000 chunks and the maximum chunk size is now 18GB only.*

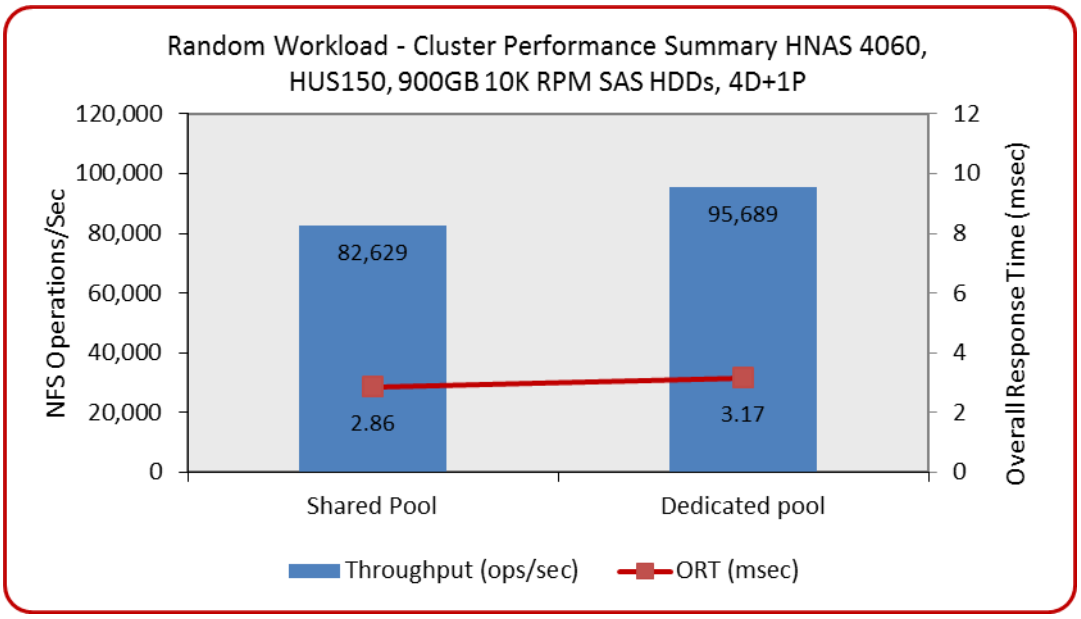
Listed below are some of the observations, recommendations and best practices one should consider before creating or expanding a storage pool in a HNAS system:

- When creating a new Storage Pool, use the same size system drives within a Storage Pool. It is strongly advised to use the same drive technology (FC/SAS/SATA) at the same spindle speed (7.2K/10K/15K RPM) within a Storage Pool.
- ***Never*** create a storage pool using just one system drive. For file system protection mechanisms, HNAS stores critical file system metadata information on multiple system drives within a storage pool. In the event of a failure or corruption, this critical data can be copied from one drive to another. But when only one system drive is used within a storage pool, there are no protection mechanisms available and conceivably puts data at risk.
- It is strongly advised to use a minimum of 2 system drives in a storage pool. When possible, use a minimum of 4 system drives in a storage pool or in each storage pool stripe set in order to maintain the minimum performance levels.
- When creating or expanding a storage pool, it is important to choose system drives from different hports, storage ports, controllers etc. That way the load on the storage pool is internally balanced

across all the available HNAS host ports, Storage ports, Storage Controllers, Front End Directors and Back End Directors etc. For more details, refer to "*Choose LUNs across different ports*" section.

- When expanding the storage pool, use the same number of system drives with the same drive characteristics. For example, if a storage pool has 4 system drives (LUNs) and if each LUN uses 15K RPM SAS drives, expand the storage pool by adding another 4 system drives, each using 15K RPM SAS Drives. This will ensure that the optimal performance levels are always maintained before and after the expansion.
- In a clustered system, when there are two or more file systems created in a storage pool, all of these file systems must be allocated to the EVS's on one node and reside on one node only. When there are multiple file systems on a storage pool and when they are assigned to two nodes within a cluster, the HNAS system automatically cuts the LUN (within that storage pool) Queue limits into half and could become performance limiting factor. Figure 12 below shows how an shared storage pool can affect the performance.
- Check and ensure that Superflush is configured on all the system drives. For more details, refer to "*Superflush*" section.
- Check and ensure that preferred path is configured for all the system drives. For more details, refer to "*sdpath*" section.

Figure 12: Negative effects of a shared storage pool across multiple cluster nodes

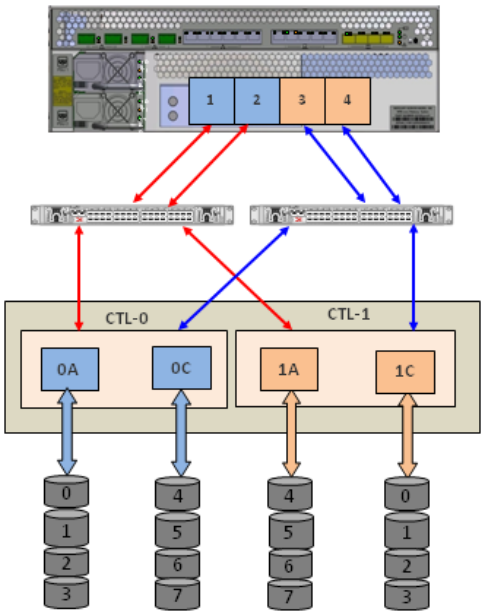


Choose LUNs across Different Ports

For a given workload, it is highly important that the load within a storage pool is internally balanced across all the available HNAS host ports, Fibre Channel Switches, Storage ports, Storage Controllers, Front End Directors, and Back End Directors etc. The easy way to achieve this is by planning ahead and ***before*** creating a storage pool.

Figure 13 below shows a simple configuration diagram using a single HNAS 3090 system and an AMS 2000 series system.

Figure 13: Configuration diagram of a simple HNAS solution



One can notice from the above figure that there are 8 LUNs mapped to the HNAS system across 4 HNAS host ports and 4 AMS target ports. Let's say the requirement is to create 2 Storage Pools on this HNAS system. *Assuming the preferred path is configured per the recommendations in the "sdpath" section*; to ensure better load distribution, achieve the optimal performance levels and eliminate any hot spots, the below LUN allocation methodology to a Storage Pool is recommended:

Table 8: Recommended LUN Layout for a Storage Pool

Storage Pool name	LUNs assigned to this Storage Pool	Why is it recommended?
SP-1	0,4,6,2	The load on each SP is distributed across the 4 HNAS hports, 4 AMS ports and the two AMS controllers
SP-2	1,5,7,3	

Table 9: Examples of poor LUN Layouts for a Storage Pool

Storage Pool name	LUNs assigned to this Storage Pool	Why they are **not** recommended?
SP-1	0,1,4,5	The load on each SP is distributed only across 2 HNAS hports, 2 AMS ports and 1 FC Switch.
SP-2	2,3,6,7	
	or	
SP-1	0,1,2,3	The load on each SP is distributed only across 2 HNAS hports and 2 AMS ports, while 4 of each are available.
SP-2	4,5,6,7	

File Systems

File system is the main storage component of an HNAS system. The maximum number of file systems in a standalone single node system is 128 and in a cluster system is 125 and can grow a maximum capacity of 256TB per file system. Listed below are some of the observations, recommendations and best practices one should consider before creating file systems in a HNAS system:

- A HNAS system supports 32KB and 4KB file system block sizes. The decision of using a 4KB or 32KB file system block size will depend on the data set characteristics, workload, performance requirements and space utilization requirements.
- A 4KB file system block size is typically more suitable for a larger number of small block I/O workloads and for a file system that will host several small files.
- A 32KB file system block size is more suitable for large block sequential workloads, large files and provides higher throughput.
- In general, the usage of 32KB file system block size is recommended if cost and capacity utilization are not a concern. 32KB file system block size tends to provide better performance for most workloads. The drawback in using 32KB file system block size is that when the average file size is small, like 4KB, a lot of disk space is wasted (a 4KB file will occupy 32KB of space). In such scenarios, the usage of 4KB block size is recommended.
- When the HNAS system is ready to destage the data to the storage layer, it first writes the data to the HNAS Sector Cache. As the HNAS sector cache is using 32KB buffers, in a write heavy system or when handling sustained write workloads, even when using a 4KB file system block size the writes are gathered into the 32KB sector cache buffers. When the Superflush is turned off or when there is low Superflush efficiency (due to fragmentation), the write IO's from the HNAS to the backend storage systems could vary between 4KB to 32KB in size. With better Superflush efficiency one can expect the HNAS to write at higher IO size (like 256KB) to the storage system.
- When using 11.1 and below releases, an individual file system can contain up to a maximum of 1023 chunks. As the file systems are based on chunks, the bigger the chunk size is, the bigger the file system can grow. For example if there are current plans to create or future plans to expand the file system to 256TB, then one should use a minimum chunk size of 320 GB or above. The usage of larger chunk size provides minor performance benefits, as the HNAS system spends less time in managing fewer chunks comparatively.
- However with the introduction of 11.2 release, a storage pool and file system can now each contain up to 60,000 chunks and the maximum chunk size is now 18GB.
- As a file system is created using a given number of chunks (allocated upfront), the following newer file systems within that storage pool are created on the chunks from where the previous file system stopped. Thus the file system that is created first resides on the outer surface of the disks and followed by the next file system which now resides on the inner surface of the disks.
- In general, one or more file systems can be created per storage pool. But the goal should be to clearly understand which file systems on that storage pool will contain active data. In the case where multiple file systems attempt to simultaneously use the storage pool that share the same system drives, disk drive seek and rotational latency may become the performance limiting factors.
- When two or more file systems are required in a storage pool, the storage pool should host the active and less frequently used file systems in order to balance the overall performance levels of that storage pool.
- In a clustered system, when there are two or more file systems created in a storage pool, all of these file systems must be allocated to the EVS's on one node and reside on one node only. When there are multiple file systems on a storage pool and when they are assigned to two nodes within a cluster, the HNAS system automatically cuts the LUN (within that storage pool) Queue limits into half and could become performance limiting factor. Figure 12 above shows how an shared storage pool can affect the performance.

- Tiered File System (**TFS**) intelligently separates the user data and Meta data onto different tiers of storage within a storage pool. When using TFS, the faster storage (from SSD, FMD or SAS drives) is assigned to Tier-0 and slower storage (from SATA, 2TB SAS) is assigned to Tier-1. As the file system metadata operations are small block random I/O's and as they are automatically placed in Tier-0, the usage of high performing disks and RAID Groups are highly recommended for Tier-0.
- While the usage of TFS would benefit certain workloads, it will not increase the peak performance limits of an HNAS system.

Superflush (HNAS release 8.2 and below)

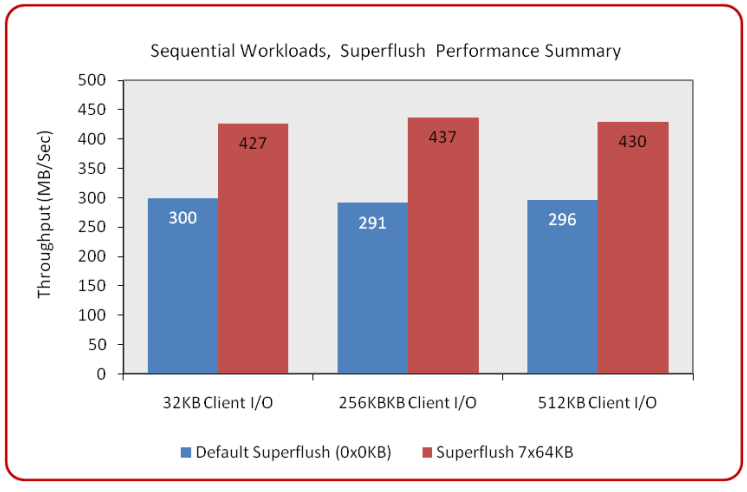
Superflush is a performance optimizing technique used by the HNAS system to maximize the efficiency of write requests sent to the storage system. Superflush is primarily efficient for system drives that are using RAID-5, RAID-6 (and to some extent, RAID-10) and has a maximum supported limit of 480KB. By attempting to write a full stripe at a time, the HNAS allows the storage system to generate parity more efficiently and improves the write/mixed workloads performance by a huge extent.

A RAID Group in the storage system is a logical mechanism that has two basic elements: a virtual block size from each disk (a chunk- incorrectly referred as stripe) and a row of chunks across the group (the RAID stripe). In the HUS/AMS family, the RAID chunk size defaults to 256KB, and is adjustable to either 64KB or 512KB (per LUN). The stripe size is the sum of the chunk sizes across a RAID Group. This only counts the “data” chunks and not any mirror or parity space. Therefore, on a RAID-5 group created as 8D+1P (nine disks), the stripe size would be 512KB (8 * 64KB chunk) or 2MB (8 * default 256KB chunk).

Irrespective of HNAS File system block size (4KB or 32KB), as the HNAS sector cache buffer size is 32KB and without the Superflush effects a single write I/O from the HNAS to the Storage system could vary between 4KB and 32KB in size. Superflush is manually configured based on the width (number of data drives in a RAID Group) and the storage chunk (stripe) size to optimize the write operations. For example, when using an 8D+1P RAID Group and the non-default AMS LUN chunk size of 64KB, the Superflush should be manually configured with a value of 8x64KB. But as this value exceeds the theoretical limits of 480KB, one should manually set the Superflush to a value closer to the theoretical limits (say, 7x64KB or 448KB).

When Superflush is configured, the HNAS attempts to coalesce/gather several write I/O's to a system drive until it reaches the configured value (in this case, 448KB) and attempts to write to the storage system using 448KB or closer I/O size. By doing so, the storage system can attempt to write at full stripe size on the LUN and thus reducing the parity related overheads. Figure 14 below shows the performance characteristics of Superflush parameter and the related performance benefits.

Figure 14: Superflush Performance Summary



Listed below are some of the observations, recommendations and best practices one should consider before configuring the Superflush parameter in a HNAS system:

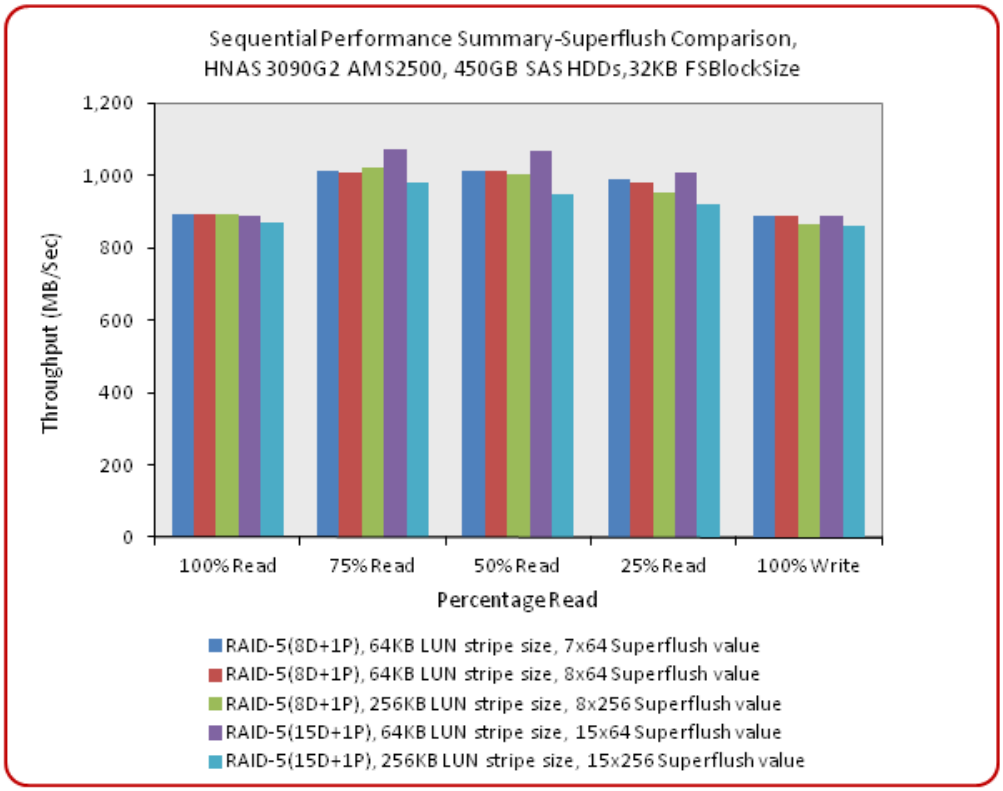
- To check the current Superflush setting, run the command “sd-list –p”. To manually configure Superflush, run the HNAS command "sd-set --width x --stripesize y <sd number>“for each system drive.
- Though theoretically, one should manually align the writes with the number of data disks in a RAID Group and the LUN Stripe Size, it is not practically possible to implement it such way on all the systems due to an upper Superflush limit of 480KB. Test results indicate that under ideal conditions the performance differences between various Superflush settings are minimal and just enabling Superflush is the key.
- When using LUNs from an HUS/AMS family storage system and default LUN Chunk (Stripe) size of 256KB, one should manually configure the Superflush value of 3 x 128. When using LUNs from an USP V, HUS VM or VSP system, one should manually configure the Superflush value of 3 x 128.
- When using LUNs from an HUS/AMS family or previous generation mid-range systems and LUN Chunk (Stripe) size of 64KB, one should manually configure the Superflush value of 7 x 64.
- Though Superflush is primarily used to optimize the write performance of the system drives using RAID-5 or RAID-6, it could assist RAID-1+0 as well.
- When creating LUNs on an HUS/AMS family storage system, one should consider formatting it using the non-default chunk size of 64KB to take full advantage of Superflush effects.
- One should also be aware that, in the event of near full capacity or aged or fragmented HNAS file systems, the chances for the HNAS system to find 480KB (maximum Superflush limits) of free contiguous blocks a time are less. During that scenario, Superflush will have no major performance benefits.
- Effective HNAS release 10.0 and above, the maximum Superflush limit has been increased from 480KB to 4MB.

Superflush (HNAS release 10.0 and above)

The concept of Superflush has not changed in HNAS software release 10.0 and above. However, the maximum Superflush limit has been increased to 4MB, meaning when using a system drive is created using RAID-5(15D+1P), 256KB LUN stripe size, one can now setup a Superflush value of 15x256.

Figure 15 below shows the performance characteristics of larger Superflush values and the related performance benefits.

Figure 15: Superflush Performance Summary



Listed below are some of the observations, recommendations and best practices one should consider before configuring larger Superflush values in a HNAS system:

- From the performance perspective, under ideal conditions the test results indicate that the differences between various Superflush values when combined with different RAID Group sizes and LUN Stripe Sizes are minimal.
- Under ideal conditions, when a HNAS system handles continuous multi client, multi files, multi threaded and heavy write workloads, the biggest IO size one can typically observe on the storage system will be up to 1MB (when using larger Superflush values like 8x256, 15x256 etc.)
- Under ideal conditions, when a HNAS system handles less writes, the biggest IO size one can typically observe on the storage system could be 1MB and above (when using larger Superflush values like 8x256, 15x256 etc.)
- However one should understand that the Superflush will be fully efficient only when the HNAS can locate the required free blocks contiguously within the file system. For example, when using RAID-5(8D+1P), 64KB LUN stripe size and 8x64 Superflush value , the HNAS should successfully locate 512KB worth of contiguous free blocks within the file system in order to attempt a single 512KB (up to) write operation.
- If the system cannot locate the required free blocks contiguously, then the Superflush will be inefficient and the HNAS could write to the storage systems using 32KB or below IO sizes.
- In the event of an aged / fragmented / near full capacity HNAS file system, the possibility of HNAS successfully locating the required contiguous free blocks is less and in such scenarios the Superflush will have very little effect.
- An optimal file system should not be filled beyond 80% capacity in order to achieve any decent Superflush efficiency.

- The combination of bigger Superflush values, larger RAID Group sizes and 256KB LUN Stripe Sizes might only be efficient under ideal conditions and in new/fresh file systems.
- It can be concluded that performance differences between various Superflush values, RAID Group and LUN Stripe Sizes are little and overtime it could only affect the overall performance levels.
- One should still consider using a smaller/moderate RAID Group sizes (say, 8D+1P), a smaller LUN stripe size (64KB) and a smaller Superflush value (say, 7x64 or 3x128) for better efficiency in the long run.

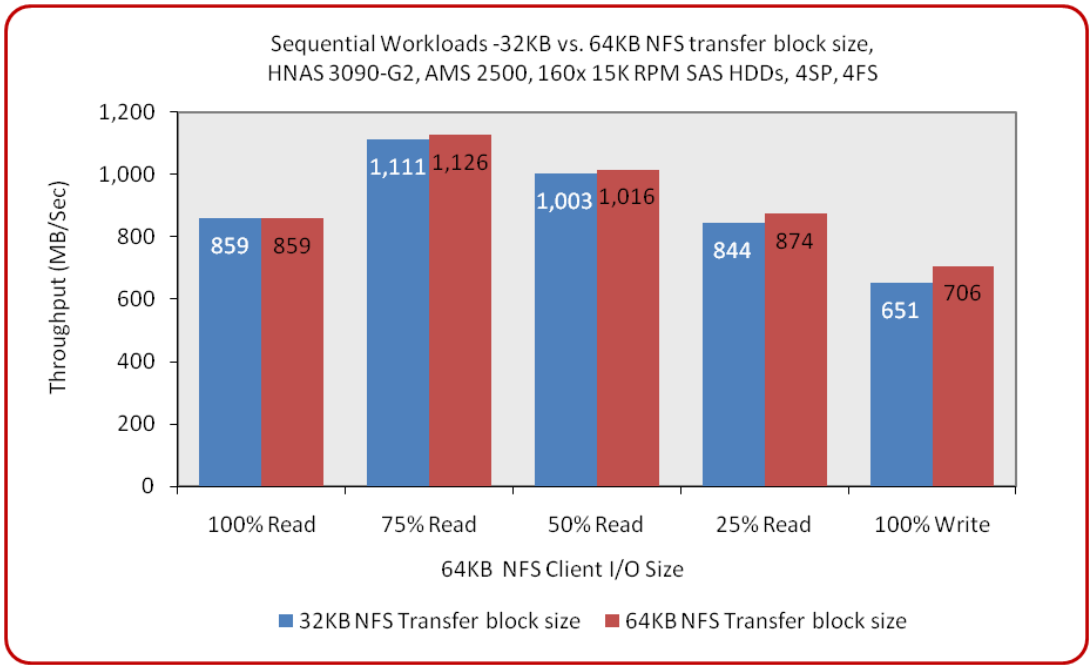
NFS

Listed below are some of the recommendations and best practices one should consider before using the HNAS system in an NFS environment:

- The HNAS system supports NFS version 2, 3 and 4. The default maximum version is 3, meaning the system supports version 2 and 3 by default. In an NFSv4 environment, one should manually change the maximum supported to 4.
- With the introduction 8.1.x release, the HNAS now supports a maximum of 64KB NFS transfer block size. One can run the HNAS commands “nfsv3-rsize-get” and “nfsv3-wsize-get” to get or set the desired values.
- The most useful NFS client performance tuning options are the ***rsize*** and ***wsize*** parameters, which defines the maximum sizes of each RPC packet for read and write operations respectively. If an HNAS file system is mounted over a high-speed network using NFS protocol, larger read and write packet sizes will enhance the overall throughput.
- When using HNAS systems running 8.1.x release or higher and when mounting a NFS export on a NFS client system, it is recommended to use the mount options "*vers=3,proto=tcp,hard,intr,timeo=600,retrans=2,wsize=65536,rsize=65536*".
- When using HNAS systems running any release lower than 8.0 and when mounting a NFS export on a NFS client system, it is recommended to use the mount options "*vers=3,proto=tcp,hard,intr,timeo=600,retrans=2,wsize=32768,rsize=32768*".
- Unlike NFSv3 protocols, the NFSv4 on the HNAS system is not fully a VLSI hardware accelerated protocol and is software based. When using NFSv4 from an HNAS system, one should *not* expect to achieve superior performance levels closer to NFSv3.

Figure 16 below shows the performance characteristics of 32KB vs. 64KB NFS transfer block sizes in an HNAS system:

Figure 16: 32KB vs. 64KB NFS transfer block size Performance Summary



iSCSI

The iSCSI protocol enables block level data transfer between requesting applications and iSCSI Target devices (HNAS). For example, using Microsoft's iSCSI Software Initiator, the Windows servers can view iSCSI targets as locally attached hard disks. Windows can then create NTFS file systems on the iSCSI targets and do I/O operations as if it were on a local disk. But on the HNAS system, the iSCSI LUNs are just regular files residing on the HNAS file systems. As a result, the use of iSCSI benefits from file system management functions provided by the HNAS system, such as NVRAM logging, snapshots and quotas.

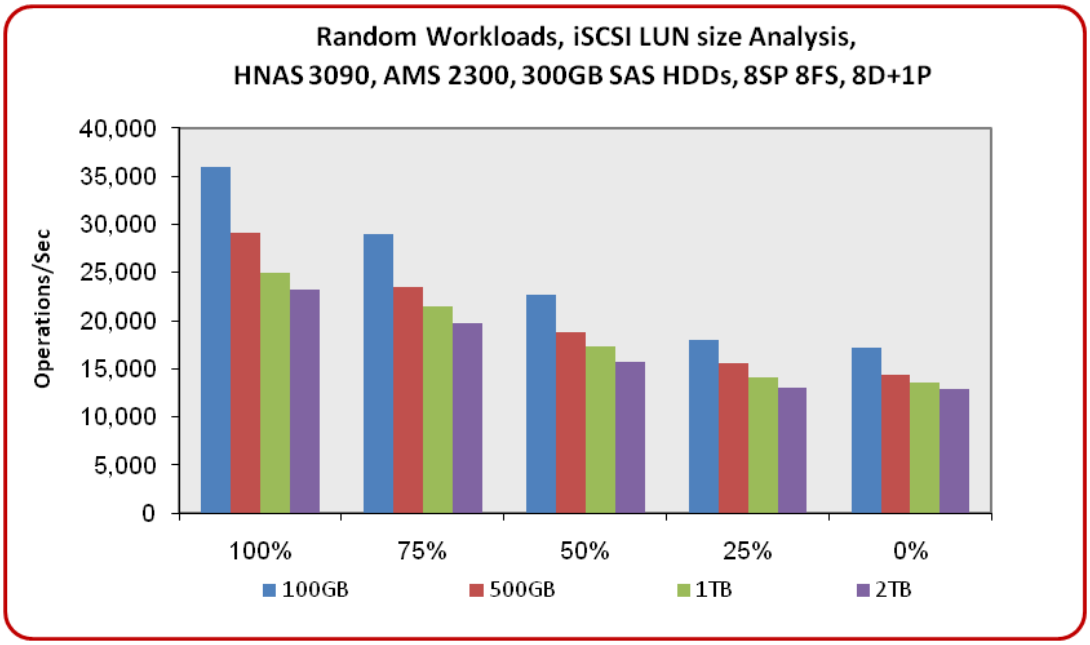
When used in Windows environment, the contents of the iSCSI LUNs are managed by the Windows client server. Where the HNAS views the iSCSI LUNs as files containing raw data, the Windows client server views each iSCSI LUN as a logical disk, and manages it as a file system volume (typically using NTFS). As a result, individual files inside of those iSCSI LUNs can only be accessed from the Windows server. Listed below are some of the recommendations and observations for using iSCSI protocol from an HNAS system:

- It is recommended that all iSCSI LUNs are placed within a well-known directory in the HNAS file system**,** for example" **/.iSCSI/**". This provides a single repository for the LUNs in a known location.
- As the iSCSI LUNs are actually files within the HNAS file systems, they can be accessed over other protocols, such as CIFS and NFS. This renders LUNs vulnerable to malicious users who can modify, rename, delete or otherwise affect them. Hence it is recommended setting sufficient security on either the LUN file, or the directory in which it resides, or both, to prevent unwanted accesses.
- It is recommended that all applications should be brought into a known state before taking a *Snapshot* of an iSCSI LUN. However, the optional software "Hitachi Application Protector" has no such restrictions and thus is recommended.

- The largest supported iSCSI LUN size from an HNAS system is 256TB. However, there's a performance penalty for using larger (like 2TB) LUN sizes. Irrespective of workloads, the overall performance levels could deteriorate when comparing the iSCSI LUN sizes from 100GB to 2TB. Past test results indicate that the optimal iSCSI LUN size is no more than 500GB.
- For both random and sequential workloads, the use of one iSCSI LUN per file system provides the best performance levels. The test results indicate that there's a performance penalty of up to 10% for hosting multiple (like 8) iSCSI LUNs per file system.

Figure 17 below shows the performance characteristics of different iSCSI LUN sizes in an HNAS system:

Figure 17: iSCSI LUN size Performance Summary



FTP

File Transfer Protocol (FTP) is a standard network protocol used to transfer files from one host to another over a TCP-based network. FTP is built on client-server architecture and utilizes separate control and data connections between the client and server. The HNAS system implements the file serving functions of an FTP server.

Listed below are some of the recommendations and best practices one should consider before using the HNAS system in an FTP environment:

- Unlike other protocols, FTP on the HNAS system is not a hardware accelerated protocol and is software based. When using FTP natively from an HNAS system, one should *not* expect to achieve higher performance levels any closer to NFSv3.
- The peak FTP throughput one could expect natively from an HNAS system is in the range of 250 - 500 MB/s.
- In an environment where higher FTP throughput is required, the usage of external FTP servers using HNAS for the data repository via NFS protocol is recommended.

Network

The HNAS 4000 systems are equipped with four integrated 10GbE ports. HNAS 3200 systems are either equipped with a NIM-3 module with two 10GbE ports or a NIM-2 module with six 1GbE ports. HNAS 3090 or 3080 systems are equipped with two 10GbE ports and six 1GbE ports. Listed below are some of the recommendations and best practices one should consider before configuring the network interfaces in the HNAS systems.

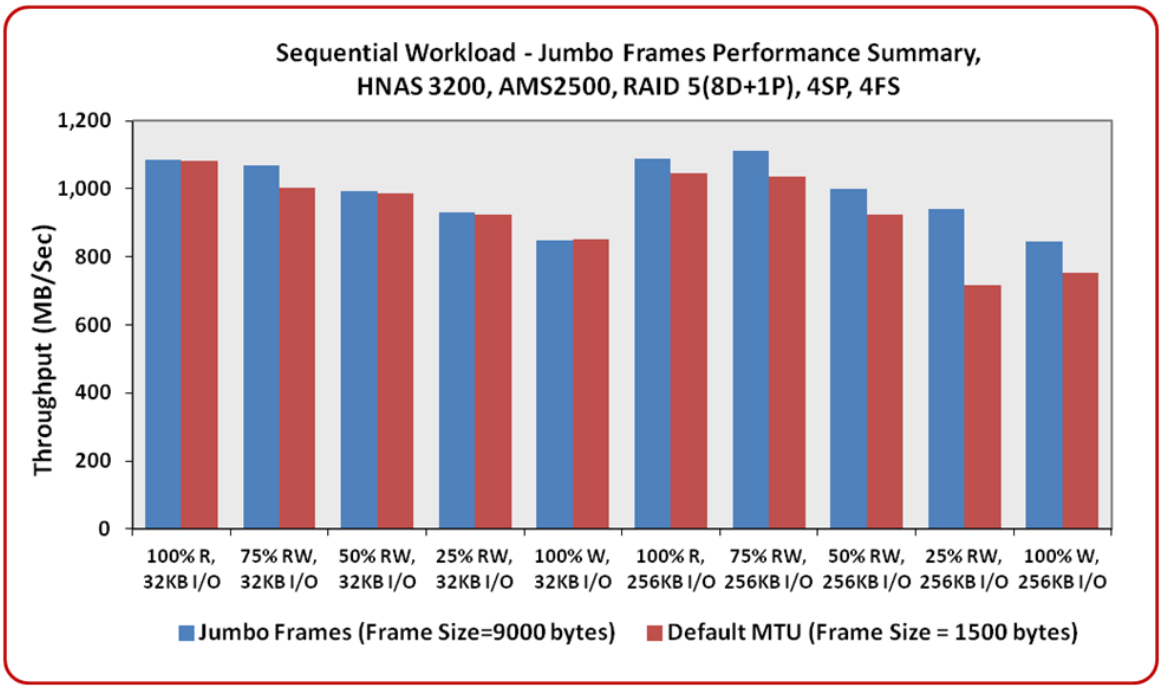
- In a 4000 system, though each of the port can deliver full duplex and dedicated bandwidth (40Gbps aggregated bandwidth across the four 10GbE ports), depending on the model the maximum achievable network throughput is only up to 2GB/s.
- In a 3200 system, though there are two 10GbE interfaces in the NIM-3 modules, they share the same transceiver and hence the available bandwidth is of just one interface. One should not, for example, configure a 3200 system with both the interfaces with the expectation that both interfaces can be driven hard simultaneously and deliver 20Gbps. The total available network bandwidth is just 10Gbps.
- In a 3090 system, though one can use all the 10GbE and 1GbE interfaces together, the maximum achievable network throughput is only up to 1.1 GB/s. (when not using the performance accelerator feature)
- In a 3080 system though one can use all the 10GbE and 1GbE interfaces together, the maximum achievable network throughput is in the range of 700 MB/s only.

Though the usage of Jumbo frames could result in performance improvement for large block sizes, the following things should be taken into consideration when planning to use jumbo frames:

- Larger frames consume more Ethernet link transmission time, causing delays for those packets that follow; thus increasing lag time and latency. This could have a negative consequence for applications that require low latency and consist of several smaller packet sizes.
- Larger frames may also fill available network switch buffer queue memory at a faster rate and may interfere with the proper operation of such gears.
- The effective use of Jumbo frames requires that every link along the network path support the same Jumbo frame MTU or sets of Jumbo frame MTUs. Without such specification erratic network behavior may result.
- Jumbo frames are more suitable for large block sequential workloads, large files and can provide higher throughput under ideal conditions.

Figure 16 below shows the performance characteristics of using default MTU vs. Jumbo frames in an HNAS system:

Figure 18: Default MTU vs. Jumbo Frames



Hitachi Performance Accelarator - Overview

One of the optional and license-based new features in the 10.0 HNAS software release is the *Hitachi Performance Accelerator* (also referred to as MaxBoost). It enables new performance features within the HNAS VLSI and provides major performance enhancements to a HNAS 3090 system. There are two components included within this feature:

1) **Throughput**: In a HNAS 3090-G2 system, there's a Fiber Channel Interface (FCI) that interfaces between the Disk Interface (DI) FPGA and the Tachyon fiber channel controller; by default it uses 4 PCIe lanes. With the Hitachi Performance Accelerator feature, it now uses all 8 PCIe lanes and thus it doubles the available bandwidth. Test results indicate that an HNAS 3090 can now deliver between 1050 to 1400 MB/sec throughput for pure sequential workloads. One should be aware that the throughput enhancement is available only when using 10.1 release and above.

2) **Ops/sec**: In this component, the number of cache controllers within the Disk Interface FPGA has been increased from 1 (default) to 2. Thus it increases the amount of cache controller processing power and resulting in about 25% higher NFSv3 ops/sec performance levels. Test results indicate that an HNAS 3090-G2 can now deliver up to 95,000 NFSv3 ops/sec under ideal conditions.

Listed below are some of the recommendations and best practices one should consider before implementing the Hitachi Performance Accelerator feature:

- The Hitachi Performance Accelerator feature is supported only on the 3090-G1 and 3090-G2 systems. The other available server models (3080, 3200 and 4000 family) do not support this feature.
- This is an optional licensed feature and can only be enabled if the related license key is present on the 3090 systems. A full system reboot (individual or cluster nodes) is also required in order to activate this feature.

- Once the license key is installed and the system is rebooted, any manual configuration related to this feature is not required.
- This feature can be uninstalled by removing the related license key and then followed by the system reboot.
- The usage of Hitachi Performance Accelerator feature will not increase the maximum performance levels of all 3090 systems. One has to understand and monitor the current performance levels of the 3090 system, available backend storage headroom (in terms of IOPS and MB/sec) and the performance requirement before installing this feature.
- If the current throughput of a 3090 system is close to or already in the range of 850-1100 MB/s, then it is likely that the usage of Hitachi Performance Accelerator feature might provide throughput improvement. Whereas if the current performance levels are not close to the peak performance levels, then this feature might not help at all.
- For configurations with a higher ops/sec requirement, one has to collect and review the HNAS Performance Information Report (PIR) before implementing this feature. Within the PIR, look for "si_busy_clocks_last_second_percentage" statistics in the logged-statistics.csv file and if that shows the SI is busy (in the range of 90-100%), and then it is likely that the usage of Hitachi Performance Accelerator feature might provide ops/sec improvement.
- Before implementing this feature, one has to make sure that sufficient backend storage headroom (in terms of storage IOPS and MB/s) is also available to handle the additional workload.

Summary

An HNAS system combined with Hitachi storage systems offer great flexibility and should perform extremely well in any number of environments. It is important that Hitachi Data Systems sales personnel, technical support staff, value added resellers and others who are responsible for the delivery of solutions fully understand the concepts, best practices and recommendations presented in this paper.

Due to the changes in the architecture and operations of these systems, all of these people need to understand why they cannot simply carry over knowledge or practices learned from other types of systems. As is true of all solutions, one must also invest the time required to design the best possible solution to meet each customer's unique requirements, whether it be capacity, availability, performance or cost. The following are the key elements to a successful solutions delivery:

- Ensure you have a clear understanding of your customer's current environment or planned environment for net new installations.
- Set the proper performance expectations. Knowing your customer's workload and clearly understanding all aspects will avoid costly mistakes in overselling the capabilities of these products.
- Poorly configured systems can lead to performance issues that could have easily been avoided with proper planning.

When in doubt, engage Hitachi Data Systems Global Solution Services (GSS) or your local reseller to assist in gathering data regarding existing storage units.

Hitachi Data Systems

Corporate Headquarters
2845 Lafayette Street
Santa Clara, California 95050-2627 USA
www.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