HITACHI
Inspire the Next

Hitachi NAS Platform Deduplication Best Practices Guide

By Francisco Salinas (Global Services Engineering)

Hitachi Data Systems

MK-92HNAS031-00

Notice

Hitachi Data Systems products and services can be ordered only under the terms and conditions of Hitachi Data Systems' applicable agreements. The use of Hitachi Data Systems products is governed by the terms of your agreements with Hitachi Data Systems.

This product includes software developed by the OpenSSL Project for use in the OpenSSL Toolkit (http://www.openssl.org/). Some parts of ADC use open source code from Network Appliance, Inc. and Traakan, Inc.

Part of the software embedded in this product is gSOAP software. Portions created by gSOAP are copyright 2001-2009 Robert A. Van Engelen, Genivia Inc. All rights reserved. The software in this product was in part provided by Genivia Inc. and any express or implied warranties, including, but not limited to, the implied warranties of merchantability and fitness for a particular purpose are disclaimed. In no event shall the author be liable for any direct, indirect, incidental, special, exemplary, or consequential damages (including, but not limited to, procurement of substitute goods or services; loss of use, data, or profits; or business interruption) however caused and on any theory of liability, whether in contract, strict liability, or tort (including negligence or otherwise) arising in any way out of the use of this software, even if advised of the possibility of such damage.

The product described in this guide may be protected by one or more U.S. patents, foreign patents, or pending applications.

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.

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.

Notice of Export Controls

Export of technical data contained in this document may require an export license from the United States government and/or the government of Japan. Contact the Hitachi Data Systems Legal Department for any export compliance questions.

Document Revision Level

Revision	Date	Description
MK-92HNAS031-00	March 2013	First publication

Contact

Hitachi Data Systems
2845 Lafayette Street
Santa Clara, California 95050-2627
https://portal.hds.com

North America: 1-800-446-0744

Purpose of this paper

The intent of this document is to define file system utilization guidelines and recommendations for end users of HDS Servers as determined by HDS Global Services.

Table of Contents

Intended audience

This guide is intended for customers and HDS field personnel. It is assumed the reader has the necessary knowledge and understanding of HNAS servers to understand the concepts discussed.

HNAS deduplication overview

HNAS 11.0 code introduces the ability to perform fixed-size block-level deduplication of a file system. The unit of deduplication is the HNAS file system block size of 4 KB or 32 KB. Deduplication (also referred to as dedupe) is post-process and occurs on a scheduled or on-demand basis. The HNAS system detects duplicate blocks within a file system by using a SHA-256 hash. Deduplication only occurs within a particular file system and not between different file systems. HNAS has dedicated FPGA state machines to perform the SHA-256 hash calculation, which reduces the impact on file serving performance and also improves dedupe performance. Two dedupe license keys are offered: basic and premium. Basic is included with all the HNAS software packages and Premium is available at an additional cost.

HNAS dedupe requirements:

- HNAS 30x0 series and newer servers
- WFS-2 file systems only
- File systems that have block-based snapshots enabled

HNAS deduplication processes

The HNAS dedupe process involves identifying whole duplicate blocks of data in the file system, removing the duplicates, and maintaining a single copy along with references. There are two types of dedupe jobs:

1. Full - A full dedupe job is only scheduled after a file system has been converted to support dedupe. During this full dedupe, the entire file system is deduped. A full dedupe can be initiated at any time.
2. Incremental - This type of job dedupes the newly added data to the file system.

When initially enabled, HNAS schedules a full dedupe on a file system.

There is one dedupe queue per node, and only one file system can be deduped at any time on a node. To make best use of the dedupe processing resources in a cluster, it is recommended to balance dedupe-enabled file systems across EVSs that reside on different nodes.

Dedupe is a read intensive process. The dedupe index resides in MMB memory, the in-memory index is 256 MB in size and is Least Recently Used (LRU) pruned. A backup of the index resides in disk.

A 32-byte SHA256 signature is calculated by one or more FPGA engines. When the process encounters two blocks with the same signature, a reference is added and the duplicate block is freed.

As data is written to a file system, HNAS tracks the amount of data written and automatically queues a dedupe job. When the job runs, HNAS takes a snapshot to index the new data and performs a dedupe job. The snapshot persists for the duration of the dedupe job.

HNAS deduplication window

The HNAS in-memory dedupe index can maintain a finite amount of unique entries. The number of entries forms the dedupe window. The benefit of this window is that irrespective of the file system size, the overhead is the same and there are no limits to the file system we can dedupe. Table 1 outlines the dedupe windows for each HNAS file system block size. The values in the table describe the amount of unique data the index can track. The table does not list the amount of data that can be deduped. The index uses an LRU algorithm to prune old entries to make room for new entries. For example, the order in which the data was written is important. In Figure 1, if duplicate data is written far enough after the original copy, there is a chance that the signatures for that data will have fallen out of the index, assuming the index became full before it reached the duplicate. If the duplicate dataset "Data1" is outside the window, it will not be deduped. To ensure the deduplication process is as effective as possible, HNAS automatically queues dedupe jobs when a certain amount of data has been written to a dedupe-enabled file system.

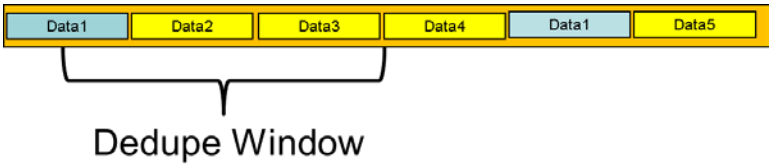


Figure 1

Table 1 lists the size of the windows based on the file system block size.

Table 1 - Dedupe window size for unique data

File system block size	
4 KB	32 KB
2.5 TB Window	2 TB Window

Table 2 lists the expected duration of a dedupe job in the best case scenario, that is, sufficient disk performance, low client load, and so forth.

Table 2 - Dedupe duration expectations

Time to dedupe (best case scenario)	25 GB	1.25 TB
Base dedupe	~5 minutes	~25 minutes
Premium dedupe	~1 minute	~5 minutes

HNAS deduplication considerations and performance

HNAS deduplication has been designed to provide an effective and minimally intrusive dedupe processing speeds. Table 3 lists the performance capabilities of dedupe for the base and the premium licenses. To maintain these dedupe throughput rates, sufficient disk resources are required.

HNAS automatically throttles back the deduplication process if user load is high to minimize the performance impact of dedupe processing. In a cluster, it is recommended to balance dedupe-enabled file systems across EVSs that reside on different nodes to make best use of the dedupe processing resources.

HNAS automatically queues a dedupe job when 1 TB of data is written to a dedupe-enabled file system. If there is a high write workload on a file system with an ingest rate higher than the dedupe speed, another job may be queued for the same file system. For example, an incremental dedupe job is queued and started. Before the dedupe jobs completes, another 1 TB of data is written. If this occurs, HNAS will abandon the current running incremental job and start a new job. HNAS will log a "too-much-change" event. This can affect the amount of data deduped on that file system. If the write activity was temporary, run a full dedupe to maximize the dedupe ratio. If the write workload is typical, higher dedupe rates can be achieved by running full dedupe jobs during periods of low write activity. For example, if an HNAS file system with a high ingest rate is being used as a backup target at night, a better dedupe rate could be accomplished by running a full dedupe job during the day for that file system.

The number of dedupe jobs extant on a node is one. If there are several dedupe-enabled file systems with high write activity, it may be better to run full dedupe jobs on these file systems to achieve the highest dedupe ratio.

Table 3 - Dedupe performance expectations

	4 KB FS Block Size	32 KB FS Block Size
Base Dedupe (1 engine)	Up to 120 MBps	Up to 120 MBps
Premium Dedupe (4 engines)	Up to 200 MBps	Up to 450 MBps

To see the dedupe throughput of the current and last dedupe jobs, issue the **fs-dedupe-history** <**fs**> command.

```
hnas:$ fs-dedupe-history fs1
Current run: none
Last run:
  Dedupe job type         : incremental
  Start time              : 2013-01-04 18:04:31-08:00
  End time                : 2013-01-04 18:05:26-08:00
  End job state           : Finished
  Data processed          : 28.2 MB
  Data deduped            : 2.35 MB
  Throughput              : 423 MB/s
```

Using HNAS deduplication

Enabling Dedupe on existing file systems

Dedupe support can be added to existing WFS-2 file systems. This is an offline process (un-mount required) using the CLI command:

fs-convert-to-support-dedupe –f <*fs_name*>

The process can take several hours for large file systems. Up to five file system can be converted at one time; however, the best practice is to convert one file system at a time. Converting more than one file system at a time may severely impact performance.

Requirements for conversion:

- File system should have sufficient free space
- No object-based snapshots
- No snapshot deletions should be taking place

The first read/write mount after conversion will automatically schedule a full dedupe run.

Note: Incremental dedupe jobs have a higher priority than full dedupe jobs. Full dedupe jobs will be paused if an incremental job is submitted to the dedupe queue.

When converting a file system, sufficient free space is required on the file system. The hash index stored on the deduped file system on HNAS 3080/3090 requires up to 45 GB of space. For a 4 K block file system, the size is 45 GB and for a 32 K block file system, it is 4.5 GB. The additional space required depends on the size of the file system. In general, it is recommended to grow the file system by one chunk and have at least 10% free space on file systems being converted. The conversion of a file system will fail if there is not sufficient free space at the end of the file system for the changes that need to be applied. The space required for the conversion will be reported in the `dblog` and `dynamic` logs if a conversion attempt fails.

Important: When converting a file system that has existing snapshots (user, backup, or replication), the size of the snapshots will increase and more free space will be required to convert the file system and run the first full dedupe job.

The conversion process can take some time depending on the capacity and block size of the file system. Refer to Table 4 for guideline conversion times.

Table 1 - Dedupe file system conversion times

	4 KB	32 KB
1 TB	4 minutes	3 minutes
10 TB	41 minutes	25 minutes
50 TB	5.25 hours	2 hours 5 minutes
100 TB	7 hours	3.5 hours

256 TB	14 hours	4 hours 15 minutes

Enabling Dedupe on new file systems

To enable dedupe on a new file system, issue the following CLI command:

format -b <block_size> --wfs 2 --dedupe-supported <*fs_name*>

By default, dedupe will be enabled. To disable, add the **--disable-dedupe** flag to the CLI format command.

A check box is available in the file system creation page on the SMU GUI to enable dedupe support for a file system.

Note: You cannot disable dedupe for a file system from the SMU GUI, this can only be achieved through the HNAS CLI.

Scheduling dedupe

The dedupe server is enabled by default and will throttle itself to minimize any performance impact on file serving.

There is no specific scheduling engine built into HNAS for dedupe. However, through cron, the dedupe service can be turned off and on. This allows one to control when deduplication runs. To turn off the dedupe service, issue the **Dedupe-service --start | --stop** command .

For example, to set up the following weekly schedule for every day of the week:

```
- Sat-Sun: Enable dedupe the whole day from 0:00 to 24:00
- Mon-Fri: Enable dedupe from 0:00 to 7:30,
           disable from 7:30 to 19:30,
           and enable again from 19:30 to 24:00
```

Use the following crontab command syntax to add a cron job:

crontab add "mm hh * * dd" "dedupe-service --start|--stop"

Where:

mm is minutes
hh is hour in 24-hour format
dd is a list of day numbers - where Sunday is **0** or **7**, Monday is **1**, Tuesday is **2**, ..., and Saturday is **6**. The list can contain one or more days separated by commas or hyphens. A hyphen is used to indicate an inclusive range. Examples of *dd* are **4** for Thursday, **1-5** for Monday through Friday,**0,6** or **6-7** for Saturday and Sunday. If *dd* is *****, it means every day of the week.

Specifically, the following three commands will support the schedule stated in our example:

```
crontab add "0 0 * * *" "dedupe-service --start"
crontab add "30 7 * * 1-5" "dedupe-service --stop"
crontab add "30 19 * * 1-5" "dedupe-service --start"
```

For more information, see examples in the `dedupe-service` man page.

Dedupe jobs can be terminated by disabling dedupe on a file system. Issue the **fs-dedupe-status-set <*fs_name*>** command. This will terminate a running dedupe job for that file system.

Note: Disabling dedupe does not inflate any previously deduped data.

Incremental dedupe jobs are scheduled when the amount of changed data in a file system crosses a predefined threshold. Incremental dedupe jobs process only the changed data and full dedupe jobs process the entire file system. Full dedupe jobs are lower priority than incremental dedupe jobs. Incrementals are triggered once a day, or when 1 TB of data changes. To manually add jobs, use the **dedupe-queue-add** command. To view the status, issue the **dedupe-status** command

Dedupe breaks snapshot size reporting (added with block-based snapshots) – deleting the snapshot may not free up the expected space.

The **df** command and the Deduplication page on the SMU report dedupe rates, as illustrated next.

```
  ID      Label     Size          Used      Snapshots       Deduped          Avail  Thin  ThinSize  ThinAvail                          FS Type
----  --------  -------  -------------  ------------  ------------  -------------  ----  --------  ---------  ---------------------------------
1024       fs1  6.82 TB  1.05 TB (15%)            NA  904 GB (46%)  5.77 TB (85%)    No                     32 KB,WFS-2,128 DSBs,dedupe enabled
```

Deduplication

Filter

File System Name: fs1

EVS:

Show: Dedupe supported

filter

Show 20 items per page

File System	Used Capacity	Reclaimed Capacity In bytes	Reclaimed Capacity Percentage	Status	EVS	Last Run
fs1	1.05 TB	903.61 GB	46%	Enabled	evs1	2013-01-06 19:22:32 (UTC-0800)

Check All | Clear All

Actions: enable disable convert

Shortcuts: File Systems Active Tasks

HNAS deduplication and virtual machine data

Virtual machine type files, such as VMware vmdk files, are ideal for deduplication. It is important to understand how HNAS handles these types of files in order to have the correct dedupe expectations. When VMware creates a Virtual Hard Disk (VHD), it stores this as a vmdk file on HNAS through NFSv3. By default, the method of allocation is thin (sparse). This is equivalent to creating a mostly empty file, which is known as a sparse file. By default, HNAS reports the full size of the file, even though it is only using a small amount of space. The HNAS dedupe engine does not dedupe the empty portions of sparse files.

For Example

- A file system contains 100 thin vmdk's of 10 GB, each deployed from the same template
  - Total space utilization reported by HNAS (by default) 1000 GBs
  - Each vmdk contains 1 GB of actual data, the rest is sparse
  - Each vmdk contains identical data
- After dedupe HNAS would report

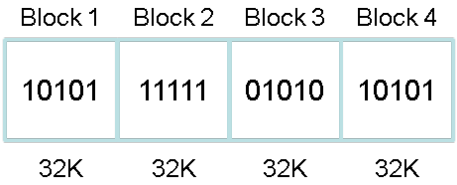
- 99 GBs deduped.
- Total utilization 901 GB

- With sparse file support enabled, HNAS would report
  - Total space utilization of 10 GB before dedupe
  - 1 GB of total space utilization after dedupe

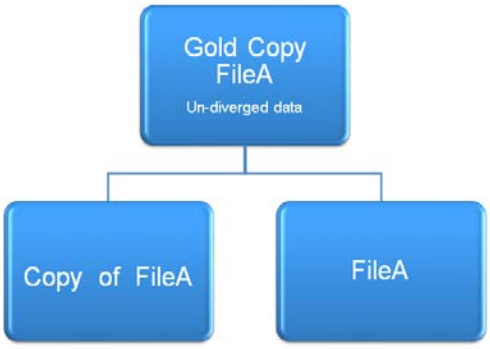
To change the way HNAS reports sparse files, use the **true-sparse-files** command. See the man page for details. True sparse files are set to “on” by default starting in HNAS release 11.1.x.

Block alignment is also important. Identical data may exist, but if the data is not aligned on the same block boundaries, HNAS will not be able to dedupe the data. This is because HNAS uses block-level deduplication.

In the next diagram, blocks 1 and 4 would be deduped, but not block 3, even though block 3 contains similar data.



HNAS file clones add another consideration for dedupe of VMware vmdk’s. File clones allow for more efficient space utilization. Dedupe treats file clones differently than regular user data. Dedupe only processes diverged blocks. When a file clone of fileA is created, FileA and Copy of FileA would contain diverged data, as illustrated in the diagram.



Deduplication interoperability

HNAS feature	Interoperability
Object replication	Data is rehydrated on replication. Source can be deduped.
File replication	Data is rehydrated on replication, the destination file system can be deduped-enabled, and the data eventually deduped
Snapshots	Dedupe is supported. ▪ Dedupe will grow any existing snapshots. ▪ File systems formatted or converted to support dedupe do not have accurate snapshot usage reporting, and the amount of space freed upon snapshot deletion can be determined. ▪ Running **kill-snapshots** on a file system that supports dedupe will leak all snapshot blocks. The **fs-reclaim-leaked-space** command will recover most of the leaked space, but **fixfs** will be needed to recover all lost blocks.
File system usage thresholds	For file systems that have dedupe support, usage thresholds can only be set through the CLI. The **fs-usage** command allows administrators to configure file system usage alert thresholds. Since snapshot usage is not available for dedupe-supported file systems, all usage is considered live. Alerts will be issued when this usage exceeds the live thresholds as well as when it exceeds the total thresholds. Alert thresholds for snapshot usage will be ignored for such file systems.
Data migrator	Dedupe can only be run on data contained within the file system. When data leaves the file system, it is rehydrated. This includes data migrator CVL and XVL. Note the following: It is possible to dedupe the secondary file system when using internal data migrator. The secondary file system must be deduped separately.
NDMP file backup	Files backed up are rehydrated to tape. When the file is restored, the full size is restored.
NDMP image backup	Data is rehydrated. Note that dedupe information may be lost on restore.
Quotas	Quota calculations are based on the length of the file (size-based quotas), or the rehydrated usage of a file (usage-based quotas).

iSCSI Logical Units	iSCSI LUs can be deduped
Tiered file system	Metadata is not deduped, only user data is deduped
File clones	Only diverged data blocks are deduped. Undivergered blocks are not. Undiverged blocks are blocks that are shared between the original and the clone.

HNAS deduplication replication considerations

HNAS supports both file and object replication with dedupe-enabled file systems. When data is replicated, any deduped data will be rehydrated.

With file replication, it is possible to separately dedupe the target file system.

Deduplication on the target file system is not supported with object replication. When demoting the original source, you cannot run any dedupe jobs against that file system; however, any data that was previously deduped will remain deduped.

Running dedupe on a file system used for replication temporarily increases the size of the file systems replication snapshots and any other snapshots on the system. Ensure the file system has sufficient free space.

Note: When running dedupe and replication concurrently on the same file system, dedupe speed will not be affect whereas replication will slow down.

HNAS Deduplication Frequently Asked Questions

See the frequently asked questions document.

Hitachi Data Systems

Corporate Headquarters
2845 Lafayette Street
Santa Clara, California 95050-2639
U.S.A.
www.hds.com

Regional Contact Information

Americas
+1 408 970 1000
info@hds.com

Europe, Middle East, and Africa
+44 (0)1753 618000
info.emea@hds.com

Asia Pacific
+852 3189 7900
hds.marketing.apac@hds.com

Hitachi Data Systems

MK-92HNAS031-00