Technical FAQ

Data ONTAP Performance Management

Clustered Data ONTAP 8.3.x

Bob Allegretti, NetApp Technical Marketing Performance Management
Version 1.2
June 2015

Abstract

This technical FAQ addresses the most commonly asked questions about performance management of the NetApp® Data ONTAP® operating system. The questions cover a broad range of performance issues that intersect with multiple NetApp technologies and features.

Version History

Version	Date	Document Version History
Version 1.1	June 2015	Updated for Data ONTAP 8.3.1 release, Bob Allegretti. Major content provided by Dimitris Krekoukias.
Version 1.0.1	April 2015	Corrections by Bob Allegretti. Moved OPM and Graphite content to separate FAQ.
Version 1.0	January 2015	Initial version for clustered Data ONTAP 8.3 by Bob Allegretti. Major content provided by Dimitris Krekoukias. Additional content and input provided by Tony Gaddis, and a multitude of individual contributors to dl-perf-ses. Large portions of Graphite and Grafana information provided by Yossi Weihs, Bryan Schramm, and Chris Madden.

TABLE OF CONTENTS

1 Overview

Will this document help me fix a performance problem?

Answer: No. This FAQ does not attempt to provide troubleshooting guidance. For performance troubleshooting issues, go to the [Performance Core Support Team (aka Perf Kx)].

What new performance improvements are in clustered Data ONTAP 8.3.1?

Answer: The primary performance improvements in 8.3.1 are with multi-core processing efficiency and All Flash FAS (AFF) latency and throughput, with adaptive inline 8KB compression enabled. The second phase of Read Fast Path to Storage read optimizations, and other modifications improve AFF latency and random read-write IOPS throughput, and enable inline compression to be used while providing comparable or better performance than Data ONTAP 8.3. More information about enhancements in Data ONTAP 8.3.1 can be found on the [Field Portal clustered Data ONTAP landing page].

How is tuning all Flash FAS (AFF) and 8.3.1 different vs previous releases?

Answer: The runtime reallocation optimizations (read_realloc, free_space_realloc) are unnecessary with AFF 8.3.1 systems, especially with smaller storage configurations (single RAID group per controller, for instance). This makes deploying AFF systems easier buy reducing configuration and optimization efforts.

However, some benefit can be realized from a regularly scheduled one-time volume (or file or LUN) reallocations on AFF systems. Schedule during off peak hours using the command line:

```
reallocate start -vserver <svm name> -force true –space-optimized -path
<path>
```

Should I enable read_realloc and free_space_realloc on all Flash FAS (AFF) systems?

Answer: See How is tuning all Flash FAS (AFF) and 8.3.1 different vs previous releases?

What new performance improvements are in clustered Data ONTAP 8.3?

Answer: There are significant performance improvements for multivolume use cases. We are continuing infrastructure work on removing single-thread bottlenecks and on balancing the utilization of multicore processors. In most cases, clustered Data ONTAP 8.3+ modestly improves performance, especially under conditions considered adverse to Data ONTAP, compared to previous versions of clustered Data ONTAP..

The greatest clustered Data ONTAP performance improvements are realized on high-core-count platforms where clustered Data ONTAP takes advantage of architectural improvements in CPU parallelism. This includes block-level deletion and overwrite improvements through batched logging and improved CIFS meta-op performance through the parallelization of CIFS lock manager. In addition, random read has improved because of read fast path (especially all-flash FAS) and write work has improved from flexible NVRAM logging (project name FlexLog). See the [Field Portal clustered Data ONTAP landing page] for more information.

Is clustered Data ONTAP faster than Data ONTAP running in 7-Mode?

Answer: Clustered Data ONTAP is feature rich in contrast to Data ONTAP running in 7-Mode. Naturally, the additional functionality comes with the cost of additional overhead. That said, clustered Data ONTAP supports faster hardware platforms that perform better. Thus, when transitioning to clustered Data ONTAP always upgrade to a platform with higher CPU clock speeds and additional core CPU counts.

What performance management features were added to clustered Data ONTAP 8.3?

Answer: Autovolumes are a QoS storage feature that provides volume-level statistics without requiring the volumes to be contained in a policy group. For the first time, clusterwide workload statistics can be collected simultaneously at the storage virtual machine (SVM) and volume level, effectively providing the functional equivalent of nested SVM and volume policy groups.

Note: Nested policy groups continue to be unsupported. For more information, see the reference to `qos statistics volume` commands in the clustered Data ONTAP 8.3 Release Notes and man pages.

For more information about using the AutoVolume CLI interface, see TR-4211: NetApp Storage Performance Primer for Clustered Data ONTAP 8.3.

Are there any known performance issues for upgrading from clustered Data ONTAP 8.2.x to 8.3.x?

Answer: No. At this time, there are no known performance issues when upgrading to clustered Data ONTAP 8.3.

Where can I go for internal discussions about performance?

Answer: Go to dl-perf-ses@netapp.com.

2 Configuration

What are the tuning recommendations for clustered Data ONTAP 8.3+?

Answer: (for spinning media and hybrid systems only, not AFF)

1. In general:

    a. Consider enabling aggregate free space reallocation with the `no_redirect` option (for more information, see the [Managing System Performance]() section of the “System Administration Guide for Cluster Administrators”). Free space reallocation continuously compacts aggregate free space during write operations resulting in longer write chain lengths and reducing reads to satisfy parity updates improving write efficiency. Plan for additional CPU overhead and disk I/O (possibly add an extra 20% headroom or pad the performance requirements by another 20%). Free space reallocation does not guarantee that file systems will **never** age over time. It is however, a preventive measure to slow the aging process and greatly reduce the frequency of scheduled aggregate and file reallocation scan operations, if required.

    Command:

    ```
    aggr modify -aggregate <aggr name> -free-space-realloc no_redirect
    ```

    b. Consider scheduling an occasional file-level (or volume or LUN) reallocation on busy data sets. However, the reallocation scan is never needed on temporary files or transaction logs because the constant over-writes nullifies any benefits realized and are a waste of resources. Reallocation scan runs in the background opportunistically when CPU resources are available. Run the reallocation scan a single time by using the `-once` option, or schedule an automated file reallocation on a volume when fragmentation reaches a critical measurement threshold (modify the `-threshold` option appropriately).

    To run it once:

    ```
    reallocate start -vserver <svm name> -space-optimized -once true -path
    <path>
    ```

    To specify a threshold of 3 (meaning schedule it only if the fragmentation threshold is over 3):

    ```
    reallocate start -vserver <svm name> -space-optimized -threshold 3
    -path <path>
    ```

    Also, if you add more RAID groups to an aggregate, NetApp strongly recommends running a forced volume reallocation for all volumes with significant I/O in that aggregate. The blocks are then redistributed across all disks:

    ```
    reallocate start -vserver <svm name> -space-optimized -force -path
    <path>
    ```

    c. Use zero-block deduplication. Clustered Data ONTAP 8.3 introduces a feature to automatically deduplicate zero user data blocks (project name `VBN_ZERO`). This feature benefits use cases in which files are initialized with all zeros, such as when VMware® provisions a new VMDK. However, for zero-block deduplication to take effect, deduplication must be enabled on the volume. Note that there is no requirement for actually scheduling any deduplication operations. You can disable deduplication operations by setting the efficiency policy to `inline-only` and setting compression to `off`. The overhead associated with zero-block deduplication is considered minimal, with potentially large performance improvements in the right use cases.

2. For database and exchange workloads with heavy sequential read after random write work, as well as disk aggregates dominated by sequential read and write workloads (which are increasingly common these days):

a. Consider enabling volume read reallocation with the space-optimized option (for more information, see the [Managing System Performance] section of the "System Administration Guide for Cluster Administrators"). Volume read reallocation and the file-level reallocation scanner described previously are very similar, bring innovative benefits, and work together in a complementary fashion. Plan on additional CPU overhead and disk I/O (possibly add an extra 20% headroom or pad the performance requirements by another 20%). Do not enable volume read reallocation on volumes that contain temporary files or transaction logs; the constant over-writes and appends on these volumes result in no benefit from read reallocation and are a waste of resources. Read reallocation is triggered by the read-ahead engine as opportunities arise to reallocate sequentially read blocks to increase read efficiency. Read reallocation does **not** guarantee in all circumstances that file systems will **never** age over time. It is, however, one preventive measure to at least slow the aging process and greatly reduce the schedule of file-level reallocation scan operations, if required.

   Command:

```
vol modify <vol name> -vserver <svm name> -read-realloc space-optimized
```

b. Consider significantly increasing the number of blocks flushed to each spindle. The default is 64 blocks x 4KB, and it can be increased on the aggregate to 256 (the setting is global for all volumes). See TR-3871, TR-4290, TR-3932, and TR-3910.

   Command (accessible from the diagnostics privilege mode):

```
aggr modify -aggregate <aggr name> max_write_alloc_blocks 256
```

3. For systems with Flash Cache, 8.3.0 introduces the option to populate Flash Cache with random writes (making a write through cache). This may help certain workloads that read recent random writes. Before enabling, measure application latency before and then after to quantify the effects.

   Command (from the node shell):

```
options flexscale.random_write_through_blocks on
```

What are some best practices to follow for optimal performance?

Answer:

- Use multiple volumes. The clustered Data ONTAP multiprocessing architecture performs best with at least 4 volumes in parallel per controller. On the largest systems consider using at least 6 volumes per node.
- On hard disk drive (HDD) aggregates, keep an adequate amount of free space (at least above 10%). As free space approaches 15%, track utilization and latency performance metrics more closely (see KB 3011413).
- To minimize latency, keep disk utilization below 50% with spinning disks. As utilization increases, so does latency. This performance factor is not WAFL specific but is a function of disk performance and queuing theory applicable to all vendors. Note this does not imply that disk utilization must always be below 50%. This applies to situations where the minimal latency possible is required.
- Keep aggregate disk counts and speeds uniform on the same controller (that is, avoid having a 100-disk aggregate and a 10-disk aggregate). The aggregate that has fewer or slower spindles and a lot of write work elongates consistency points on the entire controller and can cause a performance issue even if most of the other disks and controllers are not busy.
- Keep disk types uniform within a controller and don't mix SAS and SATA. However, NetApp Flash Pool™ is a special case.
- Remember that large file counts in a single directory can be problematic and require special sizing considerations. For more information, see TR-3537: High File-Count Environment Best Practices.

- Pace (spread out) very large file deletions.
- Keep processor utilization below 70% at all times (keeping in mind that background scanning, data protection, and opportunistic processing must be factored out). At 85% CPU utilization, internal queuing delays are most likely contributing to undesired system latency. In high-availability (HA) environments where maintaining response times is critical, keep the sum of the pair under 100% during peak business operation hours. This may seem too restrictive, but is a necessary trade-off between meeting business objectives and cost. If higher response times can be tolerated during failures with reduced capacity, the sum of the pair can be higher than 100%.
- Install the largest controller possible with memory and caching that fits the budget.
- When business objectives warrant it, trade off space efficiency for better performance and avoid using deduplication and compression.
- Spread out data management scheduled operations over time; for example, don't try to deduplicate and mirror everything at midnight.

3 Counters and Metrics

Is the `nfs-stats-top-clients` counter available in clustered Data ONTAP?

Answer: No. We will add support in a future release. There is currently no workaround.

Are the same Data ONTAP operating in 7-Mode counters, or their functional equivalents, available in clustered Data ONTAP?

Answer: No. This issue will be addressed in a future release, in the “Counter Gaps” project. Some counters have been carried over, and some have not. Until recently, there was no formal requirement for backward-compatible 7-Mode counters or a functional equivalent in clustered Data ONTAP.

What is “latency”?

Answer: Latency is the measured two-way round-trip time of an operation (as opposed to a one-way delay). It’s also known in other industries as *response time*.

What is “average processor utilization”?

Answer: This term refers to the average processor utilization across all cores in the system. As a result of improvements in clustered Data ONTAP parallelism, the reduction of single-core bottlenecks, and improvements in symmetrical processing, this metric is much more accurate now than in the past. For more details, see KB 3014084.

What is “disk utilization”?

Answer: In the context of a single disk or spindle, *disk utilization* is the percentage of time that a disk is busy processing a command. In the context of an aggregate, this term refers to the average disk utilization across all disks in an aggregate. Sometimes the term is used to reflect the highest average disk utilization across all disks in an aggregate.

4 Use Cases

When should I enable a feature?

Answer: If the controller is running at high CPU utilization during peak business operating hours or system latency is high and trending higher, do not enable any additional features. Take every action to reduce the load as much as possible. (See the FAQ Do I need to rebalance workloads?) First determine whether the feature will have the desired effect on a given workload, then fully understand the short-term and long-term impact the feature will have on the controller. For example, enabling deduplication consumes more system resources, and with certain workloads can accelerate disk aging that can cause unanticipated long-term performance problems.

Can I enable a feature?

Answer: Enable a feature only if adequate node physical resources are available to accommodate that feature for the targeted workload. You must discover the peak resource utilization during normal business operations of the node physical resources, such as average processor utilization and disk utilization.

You can collect this information during peak hours of operation by using a perfstat, CLI commands, or NetApp OnCommand® Performance Manager (OPM). You then need to estimate the overhead associated with enabling the feature. To do so, you need a solid understanding of the workload characteristics on the node and how that workload will interact with the feature (see How do I discover workload characteristics). Based on that information, you can consult feature subject-matter experts to determine whether the feature achieves the desired effect and whether the feature impact on node physical resources can be absorbed. Also read the FAQ Do I need additional hardware (nodes, disks)?

Can I add a workload to a cluster?

Answer: First you must determine whether adequate node physical resources are available to add the workload. You must discover or estimate the **workload peak resource consumption**, such as CPU and disk utilization. This resource consumption may be determined by observing similar workloads in isolation, by instrumenting the workload for monitoring (see How do I discover workload characteristics), or by testing in a staging area before deployment.

Then to determine whether the workload can be added to a node within the cluster, you must discover the **peak utilization** during normal business operations of node physical resources, such as average processor utilization and disk utilization. You must collect this information during peak hours of operation by using a perfstat, CLI commands, or OnCommand Performance Manager.

Based on the workload peak resource consumption and the node-level resource utilization, you can determine whether a new workload fits within a given node. You might come to the conclusion that additional hardware is required. Read the FAQ Do I need additional hardware (nodes, disks)?, as well. There might also be a requirement to rebalance workloads. For more information, read the FAQ Do I need to rebalance workloads?

Do I need to rebalance workloads?

Answer: Keep in mind that the primary use case for data motion through volume moves is for capacity planning and not performance. Also, if there are never performance issues, then rebalancing is optional. However, in some circumstances, you might consider proactively rebalancing for performance to avoid the possibility that a single node becomes overloaded when performance capacity is available elsewhere. Rebalancing might also be required to make room to add a workload to a cluster.

To avoid overreactions, make the effort to monitor performance over long periods and to understand resource utilization along with business operation cycles. You must discover the peak utilization during normal business operations of the physical resources, such as average processor utilization and disk utilization on every node.

A good strategy is to move volumes with the highest load between nodes within an SVM non-disruptively. Volume moves between different SVMs are disruptive. Be sure to account for the additional load incurred on the source and destination nodes from the volume move itself. Be sure to schedule the move operation during maintenance windows, off-hours, or periods of light use. For more information about data motion and volume moves, see TR-4075 and TR-4100.

To select a volume to move, you must discover the peak utilization of candidate volume workloads during normal business operations of node resources. Then predict the impact of the candidate workload at the new target node. For more information, see the FAQ Can I add a workload to a cluster? The metrics that you use to make this determination must be collected during peak hours of operation by using a perfstat, CLI commands, or OnCommand Performance Manager. Based on this information, you can make a node-based comparison to determine whether node resource utilization is skewed to a level where a rebalance is required and can be accommodated on a different node. Read the FAQ Do I need additional hardware (nodes, disks)?, as well.

Where can I move a workload?

Answer: You must understand the characteristics of the workload and find a node with physical resources available to host that workload (see How do I discover workload characteristics). This workflow is similar to the process discussed in the FAQ Can I add a workload to a cluster?

Is my storage subsystem (aggregates and HDDs) performing optimally?

Answer: If there are no complaints about latency, then the answer is nearly always yes. However, if you are concerned that latency is degrading over time and tracking with increases in disk utilization, then you might consider proactive measures to determine whether an aging file system or aggregate is potentially an issue. (**Caution: this discussion does not apply to SSDs**). Before you proceed, read and understand the following KB articles: 1010643, 2017993, 3011946, 1011510, and 3011413; and the section Managing System Performance in the "System Administration Guide for Cluster Administrators."

WAFL is designed to buffer logical write operations, rearranging disk writes into locality-optimized chains of physical disk blocks that when written to disk are striped across raid groups (called a tetris). This approach works best when the aggregate is new, with plenty of free space (if you haven't already, read KB 3011413). Over long periods of time, as data is deleted, overwritten, and added, the aggregate fills, and contiguous free space might become fragmented. As free space shrinks and becomes fragmented, the opportunities to optimize become increasingly scarce. Some workloads, such as heavy read-modify-writes, accelerate the aging process.

You must consider several pieces of information at the aggregate, RAID, and disk level to determine whether an aged file system is affecting operations. The following lists some information to consider. No single indicator taken alone is sufficient, all data must be considered as a whole:

- **Aggregate free space.** If free space is low or is fragmented, and additional indicators suggest an aged file system, then free space could be a contributing factor. Keep free space above 10%. The more free space, the better.
- **Aggregate or disk statistics,** taken from the `statit` command by using perfstat. Check the read chain length, write chain length, calculate parity reads (`cpreads`), user writes (`uwrites`), utilization (`%ut`), read latency (`ureads--*--usecs`), and write latency (`writes--*--usecs`). Small chain lengths or large variation across RAID groups within an aggregate, a high ratio of calculate parity reads to user writes, high disk utilization, and high disk latency are all indicators that suggest an aged file system. (If you haven't already, read KB 2017993.) The two types of fragmentation are free space and file. As the ratio of calculate parity reads to user writes, `cpreads/uwrites`, approaches 1, the stronger the indicator is in support of fragmented free space. In addition, when this indicator hits 1, RAID group stripes are essentially 50% full and are computationally expensive, consuming more system resources (horizontal fragmentation). Small write chain lengths also indicate limited free space with respect to a single HDD disk (vertical fragmentation). Small read chain lengths indicate file fragmentation where sequential reads are handled inefficiently.
- **RAID statistics,** taken from `statit` by using perfstat. Check the ratio of partial stripes to stripes written per second. If the ratio is high, it can be considered an indicator in support of an aged file system.

Several reallocation methods can remedy file and free space fragmentation, such as scheduling a volume file and aggregate reallocation operation from a command line or enabling features such as read reallocation and free space reallocation. Some workloads affect the file system such that a major reallocation operation on the aggregate must be executed.

The command-line option to schedule aggregate reallocation is `reallocate –A` (accessible from node shell). Aggregate reallocation affects system performance, so you must run it during a maintenance window. The time it takes to run can also be an issue on large aggregates or slow systems (on the order of days), potentially exceeding the maintenance window and interfering with business operations.

The bulk of this time is spent fixing up reallocated blocks for snapshot copies that will most likely not be accessed, referred to by reallocate as the *redirect phase* (described in more detail in BURT 768028). To avoid costly redirections, reallocate can be terminated when the redirection phase is entered, greatly reducing execution time. Reallocate can be executed once by using the `–o` option. When completed, if sufficient CPU and disk resources are available, consider the tuning recommendations that use free space reallocation and read reallocation; see What are the tuning recommendations for clustered Data ONTAP 8.3+?

If free space is the primary issue, it might be necessary to add more disks. Read the FAQ Do I need additional hardware (nodes, disks)?, as well.

Will I benefit from adding Flash Pool?

Answer: See TR-4070: NetApp Flash Pool Design and Implementation Guide, NetApp Flash Pool SE Technical Presentation, and Flash Pool Technical FAQ.

Do I need additional hardware (nodes, disks)?

Answer: You must establish safe maximum resource utilization levels and discover actual peak utilization during normal business operations. When peak resource utilization approaches the established safe maximum utilization level, more hardware might be required. Never plan on using 100%, or even close to 100%, of a resource.

You must also consider other factors such as high-availability (HA) failures and takeovers where safe maximum utilization levels could be the sum of the pair. In addition, the criticality of the workloads hosted on the resource plays a role in justifying the cost associated with lower maximum utilization levels. Establishing safe maximum utilization levels is highly specific to a customer environment. However, following is some general guidance:

- **Nodes.** Additional nodes might be required when peak CPU load is too high or peak disk utilization is too high at the maximum storage capacity for a platform. If you are concerned that peak CPU utilization is approaching safe maximum levels, it is important to know why. WAFL runs background scanners that opportunistically use available CPU cycles (such as NetApp SnapMirror® software, deduplication, deswizzle, and so on). You can manage the scanner impact by reducing the frequency or eliminating it entirely. Or you might consider the scanner impact acceptable and factor it out. However, if peak CPU load is due to normal business operations, then additional nodes might be required. Similarly, if peak disk utilization is approaching safe maximum levels due to normal business operations and it is not possible to add more disk resources, then additional nodes might be required, followed by a rebalance of workload.
- **Disks.** Providing Data ONTAP with additional HDDs can improve performance by increasing contiguous free space while providing opportunities for more I/O parallelism and throughput. The general consensus is that additional disks are not required if utilization is under 50%; additional disks are most likely required if utilization is above 85%; and additional disks might or might not be required if utilization is between 50% and 85%. (See the FAQ What are some best practices to follow for optimal performance?)

  To make an informed determination, you must fully understand both the workload characteristics and the disk access patterns. You must also fully understand application and business requirements. For example, if random read workloads are going to disk too frequently, then additional cache might be required, not additional disks. (For more information about Flash Cache, see Flash Cache Technical FAQ and TR-3832.) For another example, if free space in the aggregate is highly fragmented, reallocation might be required, not additional disks. (See the FAQ Is my storage subsystem (aggregates and HDDs) performing optimally?) After adding more disks, a file reallocation might be required to rebalance work across the aggregate. (See the FAQ What are the tuning recommendations for clustered Data ONTAP 8.3+?)

In any case, the resource utilization must be collected during peak hours of business operations by using a perfstat, CLI commands, or OnCommand Performance Manager.

Am I meeting critical service-level objectives (SLOs)?

Answer: Quantifying service levels varies among customers. One common method is to establish volume workload latency thresholds.

Note: The volume workload object is not the same as the volume storage object. (For more information, see TR-4211.)

The actual threshold values vary among customers and depend on the business criticality of the applications that use the volume. If critical workload latency approaches the SLO threshold, then you must determine the underlying cause. OnCommand Performance Manager supports threshold policies that will generate events when an SLO threshold is breached (see the OPM landing page on Field Portal). In some cases, if noncritical work is interfering, then moving the noncritical work off of a shared resource, such as a volume, node, or aggregate, can remediate the situation. In other cases, if the inherent load of the business-critical work has increased, then platform upgrades, additional disks, or additional cache might be required. (See Do I need additional hardware (nodes, disks)?) To determine whether SLOs are being met, you can collect the volume workload latency during peak hours of operation by using a perfstat, CLI commands, or OnCommand Performance Manager.

How do I discover workload characteristics?

Answer: There are two primary methods for doing this:

1. Instrument a new or observe an existing clustered Data ONTAP workload object. The workload object shows the percentage of operations that are classified sequential reads and sequential writes. From this information the random operations can be derived to complete the picture. The workload object also shows read caching statistics to help characterize the working set.
2. Use the Automated Workload Analyzer (AWA) tool that is built into clustered Data ONTAP. AWA measures the workload of an aggregate, and with 8.3.1 provide a volume-level breakdown. AWA also reports projected cache hit rates for different cache sizes, modeling the aggregate as a Flash Pool aggregate. The volume-level workload information can be used to determine whether volume placement in a HDD aggregate with Flash Cache acceleration or volume migration to an all-SSD aggregate is desirable. More information can be found in the Flash Pool Technical FAQ on Field Portal.

5 Interactions with Other Features

What are best practices regarding data protection features such as Snapshot and SnapMirror?

Answer: Stagger snap creations, deletions, and schedules. Schedule these operations during known periods of low utilization, such as off-hours. For more information, see Data Protection Top Technical Resources on the Field Portal.

What are the performance considerations with deduplication and compression?

Answer: When you enable space efficiency mechanisms such as deduplication and compression, performance is affected by the additional processing and I/O operations that occur. NetApp generally recommends enabling **compression** for use cases with data at rest (for archival purposes). Be very cautious when you consider compression for highly random workloads, especially in a heavy write environment. For more information, see TR-3966: NetApp Data Compression and Deduplication Deployment and Implementation Guide.

The All Flash FAS (AFF) systems introduced with Data ONTAP 8.3.1 have adaptive inline compression and zero-block deduplication enabled by default and are exceptions. It specifically benefits Database and Virtual Desktop Infrastructure (VDI) workloads and any other workload that does 8K I/O (or multiples of 8K). In addition, for VDI and virtualization use cases in general, Always On Deduplication is recommended on AFF with a very frequent schedule (as tight as once a minute) and CPU backgrounding to help minimize host performance impact.

When enabling deduplication, consider the following approaches to minimize the long-term impact on system performance:

1. Deduplicate datasets that are primarily at rest, not those that are highly modified on a daily basis. For example, even if you would get 99% deduplication, do **not** deduplicate a dataset that grows during the day, is deleted overnight, and then added the next day. The impact on system resources is not commensurate with the benefit.
2. Stagger deduplication operations. Scheduling everything at the default of midnight can overload the system.
3. If possible, run deduplication before snapshot copies and replications to optimize these operations. This might be difficult, or it might not be feasible to schedule on large clusters with complex data management processes. For an automated approach, see the following command:

   ```
   volume efficiency modify -vserver <SVM name> -volume <vol> -schedule
   sat@23
   ```

4. Run deduplication opportunistically by using an "auto" schedule, meaning that it runs only if 20% of data has **actually changed** in the volume. This usually auto-staggers the schedule for most customers.

   ```
   volume efficiency policy modify –vserver <SVM name> -policy
   <policyname> -type threshold –start-threshold-percent 20
   ```

5. Avoid too many deduplication streams' hitting systems in parallel by using an option to limit the number of streams (accessible from node shell):

   ```
   options sis.max_active_ops <value>
   ```

   The default is 8, but consider 1–2 for smaller systems or systems that get heavily loaded. Experiment until you reach the right balance of latency versus deduplication finishing in time.
6. Consider running deduplication in background mode:

   ```
   volume efficiency policy modify –vserver <SVM name> -policy
   <policyname> -qos-policy background
   ```

7. Consider running deduplication in performance mode (which reduces block fragmentation, it's not to make deduplication complete faster) by using the `–optimize performance` switch.

What do I need to know about high file count (HFC) environments?

Answer: These environments can lead to workloads that tend to be problematic, so it is safest to use the most powerful CPU (in terms of clock speed) and the maximum amount of memory and Flash Cache that your budget allows. For more information, see TR-3537: High File-Count Environment Best Practices. In general, it is best not to go too deep in the directory tree, but, rather, to go wide.

What are the caveats with NetApp FlexArray™ software and V-Series?

Answer: When external arrays are virtualized, Data ONTAP is unaware of the true underlying foreign array disk geometry. Thus, reallocations produce unpredictable results.

6 Tools

What tools are available for clustered Data ONTAP performance management?

Answer:

- NetApp OnCommand Performance Manager
- Perfstat Converged for Cluster Mode with LatX
- CLI commands
- NetApp AutoSupport™ diagnostics system
- wlstats
- OnCommand Insight (paid)

Also, see the Assessment Tools community portal and Counter Manager Performance Grapher (CMPG) tool.

How do I manage and monitor cluster physical resource (hardware) utilization?

Answer: Use OnCommand Performance , CLI commands, Perfstat Converged for Cluster Mode, or Graphite/Grafana. In addition:

- For more information about CLI commands for performance monitoring, see TR-4211: NetApp Storage Performance Primer for Clustered Data ONTAP 8.3.
- For more information about Perfstat Converged for Cluster Mode, see Performance and Statistics Collector at mysupport.netapp.com.

What does OnCommand Performance Manager provide?

Answer: Please see the OnCommand Performance Manager SE Technical Presentation and Technical FAQ for information.

7 Contact Us

Do you have a question that you think we should include in this FAQ? Ask us at docfeedback@netapp.com. Include “NetApp Performance Management TECHNICAL FAQ” in the subject line.

Refer to the [Interoperability Matrix Tool (IMT)]() on the NetApp Support site to validate that the exact product and feature versions described in this document are supported for your specific environment. The NetApp IMT defines the product components and versions that can be used to construct configurations that are supported by NetApp. Specific results depend on each customer's installation in accordance with published specifications.