Configuring and Tuning NetApp Storage Systems for High-Performance Random-Access Workloads

Stephen Daniel, NetApp
January 6, 2008 | TR-3647
daniel@netapp.com

Executive Summary

With the development of Data ONTAP® 7.3, NetApp continues its ongoing investments in improving the performance of its storage systems under demanding enterprise application workloads. This technical report summarizes the best practices that we have developed as a result of these investments. Most of the material in this document is applicable to both Data ONTAP 7.2 and 7.3.

1. Introduction

The development of Data ONTAP 7.3 continues to build upon the significant investments by NetApp in understanding how to achieve the highest possible performance of our storage systems when used in high-performance transactional environments. These environments tend to be characterized by workloads that are mostly small-block random access, have a significant random overwrite component to the workload, and are frequently run on storage systems dedicated to a single application or a group of related applications.

Some of these investments yielded improvements to Data ONTAP that will ship in Data ONTAP 7.3. Some of the work has led to a deeper understanding of how to set up NetApp storage systems for this style of workload. Many of these practices are applicable to both Data ONTAP 7.2 and 7.3.

2. General Best Practices for High Performance

- NetApp seeks to continually improve its Data ONTAP software. For this reason we recommend that customers seeking the best possible performance stay current with Data ONTAP releases. Early exposure of release candidates and update releases to testing environments, followed by their earliest practical adoption into production environments, usually yields the best performance.

- Random-access workloads often place significant demands on the throughput of individual disk drives. Systems must be configured with enough disk drives, even if the resulting system has more storage capacity than required by the application.

- Because these types of workloads place substantial demands on disk performance, NetApp encourages customers to configure these systems with the highest performance disk infrastructure. As of this writing, that means using 4Gbps-capable 15,000 RPM Fibre Channel disks.

- The Fibre Channel interconnect between storage controllers and disk shelves can be configured in a number of ways. The best way depends on the number of shelves, the number of available Fibre Channel ports, how the shelves will be used, whether the system is part of a cluster, and other factors. Our testing suggests that configuring the system with the maximum practical number of Fibre Channel ports and balancing the shelves across the interconnect results in a measurable difference in performance. Optimal shelf configuration may also maximize overall system availability. See TR-3437 for more information.

- NetApp offers a number of RAID technologies to protect against disk failures. For our high- performance workload studies we always recommend RAID-DP™, our performance-optimized implementation of RAID 6. We firmly believe that modern disk technology requires double-parity protection to meet today's enterprise application availability expectations.

- NetApp allows customers a choice in the size of RAID group. We recommend using the default of 16 disks per double-parity RAID group. When the number of drives in the system (after spares) is not a multiple of 16, we recommend increasing the number of drives in a RAID group up to 20 disks. Doing so ensures the most efficient rebuild times while maximizing overall capacity efficiency and system performance.

- NetApp supports managing space using FlexVol® volumes and aggregates, or using the older style volumes known as traditional volumes. For performance-oriented environments, we recommend FlexVol volumes and storage aggregates. This approach simplifies manageability by automating workload provisioning and optimization.

- Lay out high-performance data sets using a minimum number of storage aggregates. When possible, place a high-performance workload on a single aggregate to ensure optimal load balancing across the disks. However, some applications require storage from more than one aggregate for functional reasons. Also, if an application requires more space than is available in a single aggregate or more performance than is available with a single storage controller, then more than one aggregate may be used.

- NetApp storage systems require a very small amount of storage (less than 1GB) in /vol/vol0 to configure and run the storage system. We recommend placing this volume in the same aggregate as application data. This recommendation is specific to online transactional systems in which the storage supports a single application or related set of applications. In other scenarios alternative configurations may be preferred. See TR-3437 for more information.

- The Data ONTAP data layout engine, WAFL®, optimizes writes to disk to improve system performance and disk bandwidth utilization. WAFL optimization uses a small amount of free or reserve space within the aggregate. For write-intensive, high-performance workloads we recommend leaving available approximately 10% of the usable space for this optimization process. This space not only ensures high-performance writes but also functions as a buffer against unexpected demands of free space for applications that burst writes to disk.

- For NetApp systems that are SAN only we recommend setting an internal flag inside Data ONTAP to adjust the priority of the SCSI target layer. Detailed instructions for this are given in the appendix. This adjustment is appropriate for performance-sensitive SAN (FC and/or iSCSI)-only implementations. This flag is not needed for storage systems with four or more CPU cores.

- For applications that make use of large, randomly accessed data sets, it may be appropriate to make adjustments to the default memory manager policy of the NetApp storage system. The appendix shows how to select a “reuse” memory policy for volumes supporting this class of work. In general, this policy should be used when running a random-access workload with poor locality of reference. It should also be considered any time the data set is a database and the database engine’s buffer pool is larger than storage system memory. If the database engine has a small buffer pool, then using the default memory management policy will typically result in better performance.

3. Benchmark and Proof of Concept Considerations

Some features of NetApp systems are difficult to measure in a benchmark or proof of concept environment. For example, a typical NetApp customer might create Snapshot™ copies once every hour, keeping three; create them once every day, keeping two daily copies; and create them once every week, keeping two weekly copies. Measuring the performance of a system with this Snapshot schedule theoretically requires driving the system with a very realistic load for at least three weeks, until the first weekly Snapshot copy is deleted.

People desiring an understanding of NetApp performance rarely have the resources to set up a test of this magnitude. We’ve given this issue a lot of attention, and we have formulated the following recommendations for the best way to benchmark a NetApp system in which Snapshot copies are a significant portion of the workload.

- We recommend a Snapshot schedule that creates 1 Snapshot copy every 15 minutes, keeping 3, and then running the performance test for more than an hour. This gives the system time to create all three Snapshot copies and to begin deleting a well-used Snapshot copy each time a new one is created. Keeping only three Snapshot copies slightly simplifies the work the system has to do, while creating a copy every 15 minutes

adds complexity. On balance, our experiments show that this Snapshot schedule is demanding enough to be a conservative model for most real-world Snapshot schedules.

- In some benchmark or proof-of-concept tests it is not practical to fill the system with as much data as it will have during normal operation. Data ONTAP 7.3 has an internal flag that can be set to adjust how data is laid out on disk. This flag has no impact on a system once it begins to fill up. However, when the system is empty this flag can generate disk layouts that are more typical of the disk layouts used when the system is more or less full. The effect of setting this flag is small, but for some workloads this flag will generate benchmark results that are more indicative of real-world performance. The details of setting this flag are given in the appendix. Applications that perform long-term archiving of data, where data is written once and neither deleted nor overwritten, should not use this flag.

4. Final Remarks

This paper provides a number of tips and techniques for configuring NetApp systems for high performance. Most of these techniques are straightforward and well known. Using special flags to tune performance represents a benchmark-oriented compromise on our part. These flags can be used to deliver performance improvements to customers whose understanding of their workload ensures that they will use them appropriately during both the testing and deployment phases of NetApp FAS arrays. Future versions of Data ONTAP will be more self-tuning, so the flags will no longer be required.

5. Reference

- Readers of this paper should also consult TR-3437, "Storage Best Practices and Resiliency Guide," available at http://www.netapp.com/library/tr/3437.pdf.

Appendix A

This paper lists two tuning options that require the use of a setflag command. When flags are set the value is set in memory only. In order to ensure that the value is persistent, the setflag command should be added to the storage system's /etc/rc file. Specifically, these lines should be added to /etc/rc on the storage system:

```
priv set diag
setflag wafl_downgrade_target 0
setflag wafl_optimize_write_once 0
```

Note: The wafl_optimize_write_once flag is not available prior to Data ONTAP 7.3.

This paper also recommends changing the default memory management policy on some volumes. The following commands can be used to change this policy:

```
priority on
priority set enabled_components=cache
priority set volume <volume-name> cache=reuse
```

Note: The enabled components subcommand is not available prior to Data ONTAP 7.3.