Technical Report

NetApp SolidFire Quality of Service (QoS)

Alexander Sammer, NetApp
October 2017 | TR-4644

Abstract

NetApp® SolidFire® Quality of Service (QoS) enables unprecedented control over performance in large-scale multitenant environments. By setting the QoS levels on the volumes in a multitenant environment, customers can guarantee performance SLAs. To guarantee the performance of a volume, it is necessary to understand what factors affect IOPS and how system performance can be guaranteed in different environments with varying workloads.

NetApp®

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 Introduction

The increasing demands and velocity of change imposed on enterprise IT organizations and service providers today is unrelenting. Users want to deploy more applications faster, and the resources that support those applications must be more agile and adapt to changing demands. Even more challenging for IT, the world's largest public cloud providers have set the benchmark for delivering infrastructure as a service (IaaS) at scale. So IT organizations are looking for infrastructure solutions capable of delivering compute, networking, and storage predictably and on demand. These solutions must enable them to dramatically raise operational efficiencies, innovate more quickly, and respond to application and business challenges faster than ever before.

At the heart of delivering infrastructure on demand, and as a service, is the concept of multitenancy, in which multiple applications or customers use the same storage infrastructure. Although the opportunity to run a broad array of applications within a single system might sound appealing, the reality for today's IT managers is very different. When many performance-sensitive applications are consolidated onto a single platform (traditional or flash), "noisy neighbor" applications cause resource contention, unpredictable application performance, and unhappy customers.

1.1 Introduction to QoS

There is a large imbalance today between the performance and capacity resources of traditional storage systems. Capacity is plentiful and low cost; conversely, input/output operations per second (IOPS) are scarce and very expensive. From a provisioning perspective, performance and capacity are rigidly bound together, which only makes matters worse. This bind forces administrators to unnecessarily add storage capacity to increase the amount of IOPS available to a particular application. What results is a wasteful allocation of resources in an effort to overcome the limitations of existing storage architectures. For a long time, the promise of delivering storage resources predictably to a broad set of applications without worry was nothing more than a pipe dream.

But now, service providers and enterprise IT can turn to the NetApp® SolidFire® storage operating system. From its inception, one of the primary problems SolidFire was designed to solve was how to effectively implement and manage quality of service (QoS). Within SolidFire, each volume is configured with minimum, maximum, and burst IOPS values. The minimum IOPS setting provides a guarantee for performance, independent of what other applications on the system are doing. The maximum and burst values control the allocation of performance so that the system delivers consistent performance to all workloads. For enterprise and service providers, SolidFire QoS enables SLAs for minimum performance metrics, providing a positive customer experience. For infrastructure consumers, hard QoS delivers clear expectations for storage performance and the ability to deploy all tier 1 and tier 2 applications with confidence.

In addition to individual volume control, administrators can deploy QoS policies that simplify operation by grouping volumes for easy management and modifications. Administrators can apply a set of policies against multiple volumes based on their performance needs.

1.2 The Impact of QoS in the Enterprise

Enterprises today are tasked with figuring out how to build a flexible, scalable platform that can support multiple workloads while improving operational efficiency. Until now, storage administrators spent the bulk of their time tuning, tweaking, planning, and troubleshooting storage performance. They face several difficulties:

- Identifying and protecting applications that have different I/O patterns
- Managing separate, siloed storage appliances, each corresponding to a separate workload
- Overcoming the difficulty of sizing storage for both initial workload placement and growth over time
- Eliminating inefficiencies and waste in capacity, performance, and operational management

It doesn't have to be this way. NetApp SolidFire provides predictable, flexible, and easily managed storage. SolidFire QoS is part of an overall virtualized system that offers the performance and availability that today's end-user workloads demand.

What enterprises need is more flexibility, and a solution that allows them to provision capacity and performance separately and uniquely for every application, every time. QoS gives them this ability.

One of the most effective ways for enterprise customers to take advantage of QoS is by consolidating multiple workloads, typically ones that were previously isolated in separate storage silos. By permitting many applications to be deployed to a single platform with guaranteed QoS, enterprise IT can now easily address all performance-related challenges within a single storage system. Enterprises can realize the following benefits:

- By reducing the number of storage platforms and vendors in use, the cost of operations goes down, and the number of tools needed to manage storage decreases.
- By provisioning capacity and performance separately out of a single pool, enterprises can do less overprovisioning to meet the needs of the workloads.
- By providing a scalable platform that can grow or shrink according to the needs of the business, enterprises can make more efficient use of capital, space and power, and labor.

Nowhere is this concept of consolidation more powerful or relevant than in the virtualized infrastructures of today's enterprise IT. The ability to provision capacity and performance separately from a storage platform unifies the resource management processes in a way that wasn't possible before. This approach is similar to the way enterprises have been using cloud management systems to provision CPU and RAM separately for years.

2 SolidFire QoS

2.1 Understanding the Basic QoS Requirements

In setting QoS levels, there are four basic contribution factors to consider:

- Desired level of performance
- Tolerance of variable performance
- Consistency of desired performance
- Value of providing this level of performance

NetApp SolidFire QoS technology enables IT administrators to easily assign and guarantee levels of performance (IOPS and bandwidth) to thousands of volumes residing in a single storage platform.

This approach proactively provides applications with the performance they require throughout the life of their deployment. With guaranteed QoS from SolidFire, applications no longer compete for performance, and administrators no longer must struggle with complex tiering systems or prioritization schemes. You allocate each storage volume capacity and performance, which you can change dynamically without migrating data or affecting performance. The performance technology is enabled through several key functions:

- **All-SSD storage.** Enables delivery of fast, consistent, and predictable I/O regardless of where the data is placed on the solid-state drive (SSD).
- **Scale-out architecture.** Linear, predictable performance building block as the system scales out.
- **RAID-less data protection.** No additional protection overhead and p**r**edictable performance in any failure condition.
- **Balanced load distribution.** Data is evenly distributed on the system, **e**liminating hot spots that create unpredictable I/O latency.

- **Fine-grain QoS control.** Set minimum, maximum and burst limits to eliminate noisy neighbors and guarantee volume performance. The minimum IOPS guarantee prevents performance headaches.
- **Performance virtualization.** Control performance on demand and independent of capacity.

Allocate, Manage, and Guarantee Storage Performance

Volumes provisioned within a SolidFire system are assigned three performance values: Min IOPS, Max IOPS, and Burst IOPS. Each value can be monitored, tracked for chargeback, and changed dynamically without affecting volume or system performance.

- **Min IOPS.** IOPS that are always available to the volume. The Min IOPS value guarantees performance regardless of system condition or application activity.

  The Min IOPS value is the minimum number of IOPS that an administrator grants to a volume. This IOPS level is what is effectively "guaranteed" and is the focus of most conservative SLA provisions. Min IOPS values come into play only if the system becomes bound by I/O capacity. At that point, the system scales all volumes back from their Max IOPS level proportionally toward their Min IOPS values. This ensures fair resource allocation when the system is heavily loaded and offers a way to prioritize more important volumes, while others are scaled back more dramatically.
- **Max IOPS.** The maximum IOPS that a volume can process over a sustained period.

  Applications are not permitted to consistently exceed this level and affect other applications. This prevents volumes from affecting a system beyond the set limits—for example, during boot storms.
- **Burst IOPS.** IOPS that a volume can process during a spike in demand. This value is particularly effective for uneven and latency-sensitive workloads.

  When a volume uses less than its Max IOPS, it accumulates credits, which can be used to burst to a volume's Burst IOPS limit for a short period. A Burst IOPS value is particularly effective for virtual machine reboots, migrations, large file transfers, and other heavy loads that need to be completed within a short period. This functionality is allowed only when system performance resources are available, preventing any impact on other applications.

As QoS becomes a must-have component of a storage infrastructure, the differences between QoS features and a purpose-built QoS architecture become evident. The SolidFire all-flash storage system is purpose-built to enable IT organizations to allocate, manage, and guarantee storage performance—making it faster and easier to respond to changing demands of applications and the business.

3 Provisioning IOPS for QoS

The following use cases illustrate what factors affect IOPS and how system performance can be guaranteed in different environments while operating under varying workloads.

3.1 QoS Use Cases

Database Use Case

Typical databases transfer with 4KB and 8KB block sizes. This example shows a medium-scale database that desires about 3000 IOPS. Lower-scale databases are in the range of 500–1000 IOPS and higher-scale are above 6000 IOPS. Databases are typically sensitive to fluctuations in performance, so the Min IOPS and the Max IOPS settings are very close together. The burst is configured only slightly higher than the maximum because databases typically push consistent traffic. Based on the performance requirements, this tier is charged the highest amount per IOPS.

Table 1) Database example.

Desired SolidFire IOPS	Typical Queue Depth	SolidFire Min IOPS	SolidFire Max IOPS	SolidFire Burst IOPS
3000	16	3000	4000	5000

Application Use Case

This tier is probably the most variable and depends on the application. This example assumes 50 users using the application, each user requiring an average of 20 IOPS. Applications typically require a relatively consistent amount of performance; however, cost is also a significant factor, so separation between Min IOPS and Max IOPS is wider.

Table 2) Application example.

Desired SolidFire IOPS	Typical Queue Depth	SolidFire Min IOPS	SolidFire Max IOPS	SolidFire Burst IOPS
1000	8	750	1250	1500

Web Server Use Case

Based on the assumption of a 1MB average page size, the following settings are for a web server that has approximately 10,000 page views per hour. Depending on the number of estimated page views per hour, this number can be scaled up or down.

The high burst value is due to the potential spike in web traffic that is associated with web servers.

Table 3) Web server example.

Desired SolidFire IOPS	Typical Queue Depth	SolidFire Min IOPS	SolidFire Max IOPS	SolidFire Burst IOPS
400	4	300	500	2000

Development Use Case

Development instances typically push a low average performance. In this instance, the lower cost is key, and they tolerate inconsistencies in performance, so min and max are widely separated. Finally, development instances have the highest levels of performance variation. In this environment, large file transfers and software builds can cause usage to spike.

Table 4) Development example.

Desired SolidFire IOPS	Typical Queue Depth	SolidFire Min IOPS	SolidFire Max IOPS	SolidFire Burst IOPS
300	4	100	500	4000

VDI Use Case

Virtual desktop infrastructure (VDI) instances (typically one VDI user refers to one instance) typically use 8–12 IOPS; however, a high percentage of reads consist of larger block sizes. Due to these larger block sizes, typical instances use 100–150 4KB IOPS. These instances want low performance with the lowest

cost. From a selling perspective, the highest possible density is key. The burst value is set to 500 because most virtual machine reboots require 400–500 SolidFire IOPS in a controlled fashion.

Table 5) VDI example.

Desired SolidFire IOPS	Typical Queue Depth	SolidFire Min IOPS	SolidFire Max IOPS	SolidFire Burst IOPS
100–150	1	100	200	500

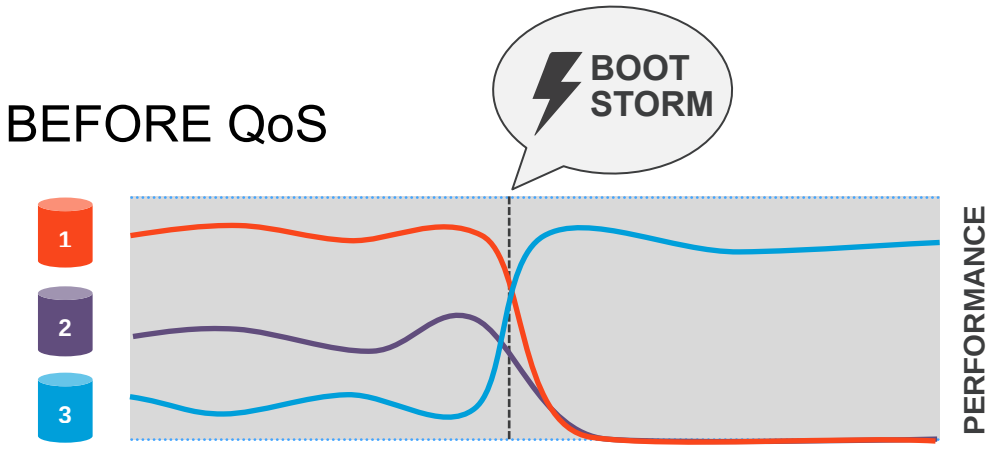
Additional Considerations

To ensure that SolidFire QoS directly controls the instances' performance, a 1:1 instance-to-volume ratio must be used. However, in practice, a 1:1 ratio might not be practical because of the 256-volume limitations of certain hypervisors. In these cases, place lower performing instances used by the same customer in the same volume. To get an accurate QoS value for the volume, sum the QoS values. For example, if a customer has 5 VDI instances, they can be combined into the same volume with a Min IOPS of 500, Max IOPS of 1000, and Burst IOPS of 2500. To prevent noisy neighbor effects within the volume, NetApp recommends no more than 16 instances per volume

Typical SLAs are written around the SolidFire QoS Min IOPS setting; however, additional agreements can be made around Max IOPS and Burst IOPS. The following section gives the two options for provisioning the cluster and guaranteeing a performance SLA.

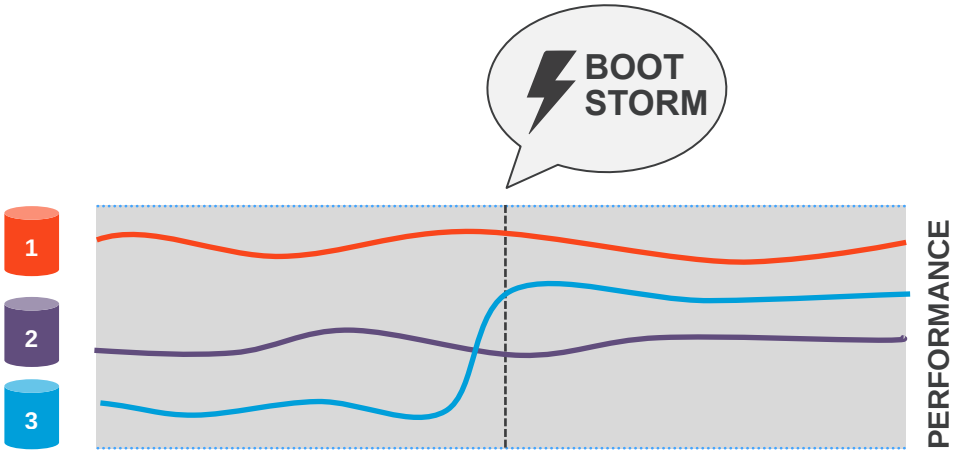
The ability to set minimum QoS values helps avoid performance impact during VDI boot storms. If you don't set a minimum QoS, the performance of all other applications decreases significantly during a VDI boot storm, as shown in Figure 1. In the graph, volume 3 is running the VDI application.

Figure 1) Performance impact during a VDI boot storm.



After a minimum QoS performance policy is set, the other applications are not affected by the VDI boot storm. Operations can continue with the same performance, as shown in Figure 2.

Figure 2) Performance after the implementation of guaranteed QoS.



3.2 Cluster Provisioning and Guaranteeing a Performance SLA

The options for cluster provisioning and guaranteeing a performance SLA are increased queue depth and fixed queue depth.

Increased Queue Depth

This option is the easiest to provision and implement. The main assumption is that applications that require higher performance use a higher queue depth to service their requests. With this method, you can utilize the full cluster performance while pushing high-performance volumes. You can use one of the following end-to-end latency values, depending on the environment.

Table 6) Latency-oriented environments.

Low Load Environment		Highly Utilized Latency Optimized Environment	
IOPS	Required Queue Depth	IOPS	Required Queue Depth
100–999	1	100–399	1
1000–1999	2	400–799	2
2000–3999	4	800–1599	4
4000–7999	8	1600–3199	8
8000+	16	3200–6399	16
		6400–12799	32
		12800+	64*

* Instances that require a queue depth above 32 must use multiple iSCSI sessions.

Table 7) Oversubscribed and highly utilized environments.

Medium Environment		Oversubscribed and Highly Utilized Environments	
IOPS	**Required Queue Depth**	**IOPS**	**Required Queue Depth**
100–999	1	100–133	1
200–399	2	134–266	2
400–799	4	267–533	4
800–1599	8	534–1066	8
1600–3199	16	1067–2133	16
3200–6399	32	2134–4266	32
6400–12799	64*	4267–8533	64*
12800+	128*	8534+	128*

*For instances that require a queue depth above 32, multiple iSCSI sessions must be used.

For more information about changing queue depth, refer to the Windows and Linux configuration guides:

Fixed Queue Depth

Certain applications require a low queue depth while still maintaining a high level of performance. With this option, Min IOPS values can be obtained for these applications; however, because of the insufficient clusterwide queue depth, full cluster performance is not likely to be reached. Another consideration is that the amount of load on the end-to-end system that various workloads create will vary. The SolidFire QoS curve takes into account the varying block sizes sent to the array to ensure that the load to the array remains the same regardless of block size. Typically, hypervisors load more quickly with many small block sizes rather than a few large block sizes. In addition, read versus write percentages load the end-to-end system differently. However, in most environments, the law of averages applies because of the thousands of volumes and applications being used.

We tested the following use cases and verified that the Min IOPS was obtained. The workload consisted of a conservative read/write approach (50% read, 50% write) and real-world block sizes. Depending on the end-to-end environment, results might vary.

Table 8) Typical configuration example.

Instance	Price	Count	Typical Queue Depth	Min IOPS	Max IOPS	Burst IOPS
Database	$$$$	20	16	3000	4000	5000
Application	$$$	50	8	750	1250	1500
Web server	$$	200	4	300	500	2000
Development	$$	50	4	100	500	4000
VDI	$	300	1	100	200	500
Total		**625**		**192,500**	**347,500**	**950,000**

In this example, there is a mixture of low-performance volumes with higher-performance volumes. This mixture typically provides the highest provisioning from both a performance and capacity standpoint. The

lower-performing volumes use more capacity with less performance, whereas the higher-performance volumes use more IOPS and less capacity.

Volume Utilization

In the previous examples, all volumes were pushed with 100% utilization for the full duration, and they were able to maintain their Min IOPS values. In typical environments, volumes push significantly less than 100% utilization—usually only a third. Most users don't use most of their allocated performance, so, with correct cluster performance monitoring and SLA flexibility, overprovisioning the Min IOPS can be advantageous.

The single volume utilization statistic can determine which volumes are using a high percentage of their allocated performance and increasing their QoS values to meet their demand.

Cluster Performance–Altering Scenarios

On a SolidFire array, there are four types of performance-altering scenarios to account for, which depend on the structure of the SLA.

- During drive failures and rebuilds, the cluster operates at 80% performance. Typical drive rebuilds take less than 10 minutes.
- During node failures and rebuilds, the cluster operates at 50% performance. Typical node rebuilds take less than 60 minutes.
- During cluster firmware upgrades, nodes might reboot to apply the new software. During the reboot period, the proportional number of performance is unavailable. For example, on a 10-node cluster, during the node reboot, the cluster operates at 90% of performance.
- During node additions, the cluster operates at 50–80% of performance, depending on the cluster size and number of simultaneous node additions.

QoS Based on Cluster Performance

- QoS settings are not reservation-based. They permit oversubscription. The user needs to ensure that the minimum QoS does not exceed the maximum IOPS for the cluster.
- The SolidFire cluster monitors the internal load to dynamically adjust the maximum.
- The total system load (measured at the node level) determines the target IOPS:
  - 0%–38%: target IOPS = Max IOPS
  - 38%–60%: target IOPS decrease linearly from Max IOPS to Min IOPS
  - 60%–100%: target IOPS decrease linearly from Min IOPS to 0

Note: The system pushes back proportionally on all volumes according to QoS settings.

SolidFire QoS guarantees an unprecedented level of control over performance through Min IOPS, Max IOPS, and Burst IOPS. Through correct provisioning, offerings from high-performance databases to low-performance VDI can be provided within the same storage array. The objective of SolidFire QoS is to maintain this precise level of provisioned performance independent of other applications running on the same multitenant array.

4 Configuring SolidFire QoS

4.1 Setting Performance SLAs

Although IOPS settings and enforcement are the basis of ideal QoS capability, administrators should consider additional operational factors when setting performance expectations and related SLAs for

internal (enterprise) and external (service provider) customers. When developing performance-based SLAs, consider the four key areas that follow.

System Provisioning

At the heart of any SLA strategy is a stance on storage volume provisioning. An aggressive provisioning (that is, heavily oversubscribed) strategy dictates a more conservative SLA strategy. However, if you provision in a very conservative manner, there is room to take a more aggressive stance with SLAs.

Understanding Total System Load

When crafting a SLA, consider factors beyond the purely quantitative IOPS metrics. An IOPS-centric approach fails to fully capture the impact of varying block sizes on overall performance. Accounting for this impact requires a more comprehensive approach—one that considers system load. System load is a function of IOPS and average block size. Incorporating these two variables into a holistic metric produces a more accurate indication of the actual load on the system.

Load Balancing

In cloud block storage, the most common workload involves large amounts of small, random I/O spread across numerous application volumes. However, it is also critical to account for the potential performance impact of other workload profiles (for example, in an instance where there is high concentration of I/O into a few volumes). Well-written SLAs create awareness and establish appropriate expectations for these outliers.

Impact of Failure Conditions

When writing performance-based SLAs, contemplate the impact of component-level (for example, disk) failure on both capacity and system-level performance. One way to account for potential performance degradation under failure conditions is to create a performance-level guarantee within the SLA, which commits to a specified level of performance for a percentage of time (for example, 95%). The buffer outside this performance level (for example, 5%) leaves appropriate room to absorb performance degradation under certain failure conditions.

4.2 Defining QoS

Understanding QoS

A NetApp SolidFire cluster can provide QoS parameters on a per-volume basis.

Cluster performance is measured in IOPS. Three configurable parameters define QoS: Min IOPS, Max IOPS, and Burst IOPS. For minimum, maximum, and default QoS values, see Table 9. IOPS parameters are defined in the following ways:

- **Min IOPS.** The Min IOPS configured for a volume is the guaranteed level of performance for a volume. Performance does not drop below this level.
- **Max IOPS.** The maximum number of sustained IOPS that the SolidFire cluster provides to a volume.
- **Burst IOPS.** The maximum number of IOPS allowed in a short-burst scenario.

  SolidFire uses Burst IOPS when a cluster is running in a state of low cluster IOPS utilization. A single volume can accrue Burst IOPS and use the credits to burst above its Max IOPS up to its Burst IOPS level for a set “burst period.” A volume can burst for up to 60 seconds if the cluster has the capacity to accommodate the burst.

  A volume accrues 1 second of burst credit (up to a maximum of 60 seconds) for every second that the volume runs below its Max IOPS limit.

Burst IOPS are limited in two ways:

- A volume can burst above its Max IOPS for a few seconds equal to the number of burst credits that the volume has accrued. After exhausting all burst credits, the volume performance is limited by its Max IOPS setting.
- When a volume bursts above its Max IOPS setting, it is limited by its Burst IOPS setting. Therefore, the Burst IOPS never exceed the Burst IOPS setting for the volume.

- **Effective Max Bandwidth.** The maximum bandwidth is calculated by multiplying the number of IOPS (based on the QoS curve) by the I/O size.

Note: QoS parameter settings of 100 Min IOPS, 1000 Max IOPS, and 1500 Burst IOPS have the following effects on quality of performance:

- Workloads are able to reach and sustain a maximum of 1000 IOPS until the condition of workload contention for IOPS becomes apparent on the cluster. IOPS are then reduced incrementally until IOPS on all volumes are within the designated QoS ranges and contention for performance is relieved.
- Performance on all volumes is pushed toward the Min IOPS of 100. Levels do not drop below the Min IOPS setting but could remain higher than 100 IOPS when workload contention is relieved.
- Performance is never greater than 1000 IOPS, or less than 100 IOPS for a sustained period. Performance of 1500 IOPS (Burst IOPS) is allowed for a short period, but only for those volumes that have accrued burst credits by running below Max IOPS. Burst levels are never sustained.

Table 9) Possible minimum and maximum values for QoS.

			I/O Size Max			
Parameters	**Min Allowed**	**Default**	**4KB**	**8KB**	**16KB**	**256KB**
Min IOPS	50	100	15,000	9375	5556	500
Max IOPS	100	15,000	200,000	--	--	--
Burst IOPS	100	15,000	200,000	62,500	37,037	333

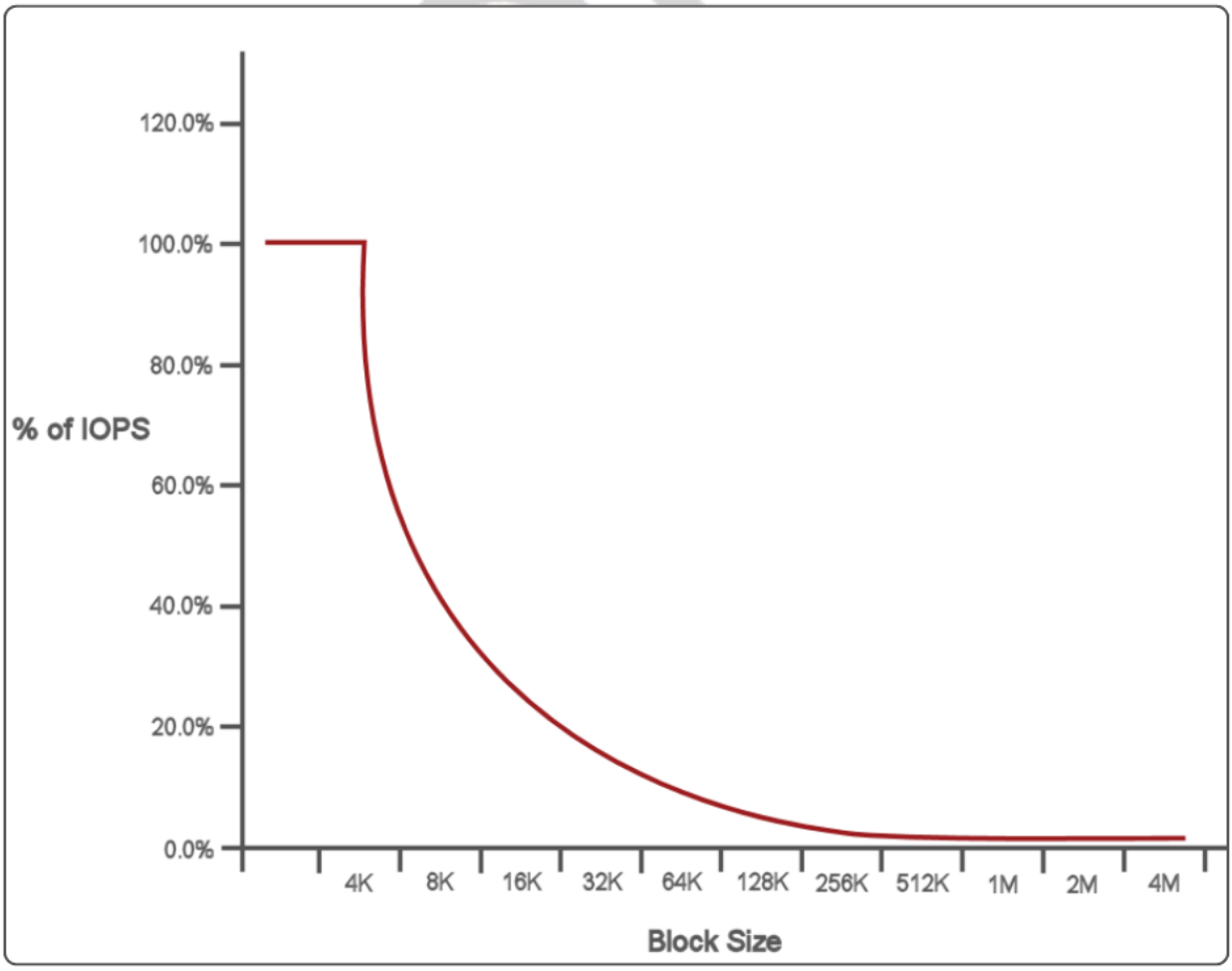
Note: Max IOPS and Burst IOPS can be set as high as 200,000; however, this setting is allowed only to effectively “uncap” the performance of a volume. Real-world maximum performance of a volume is limited by cluster usage and per-node performance.

QoS Performance Curve

Block size and bandwidth have a direct impact on the number of IOPS that an application can obtain. SolidFire software takes into account the block sizes it receives by normalizing block sizes to 4KB. Based on workload, the system might increase block sizes. As block sizes increase, the system increases

bandwidth to a level necessary to process the larger block sizes. As bandwidth increases, the number of IOPS the system can attain decreases.

Figure 3) QoS performance curve.



For example, if block sizes are 4KB, and bandwidth is 4000KBps, the IOPS are 1000. If block sizes increase to 8KB, bandwidth increases to 5000KBps, and IOPS decrease to 625. By taking block size into account, the system ensures that lower-priority workloads that use higher block sizes, such as backups and hypervisor activities, do not take too much of the performance needed by higher-priority traffic using smaller block sizes.

4.3 Configuring QoS

Configuring QoS for Volumes

You can specify QoS settings for a volume during creation, either through the GUI or through SolidFire API calls. For more information about the API, refer to the user documentation listed at the end of this report.

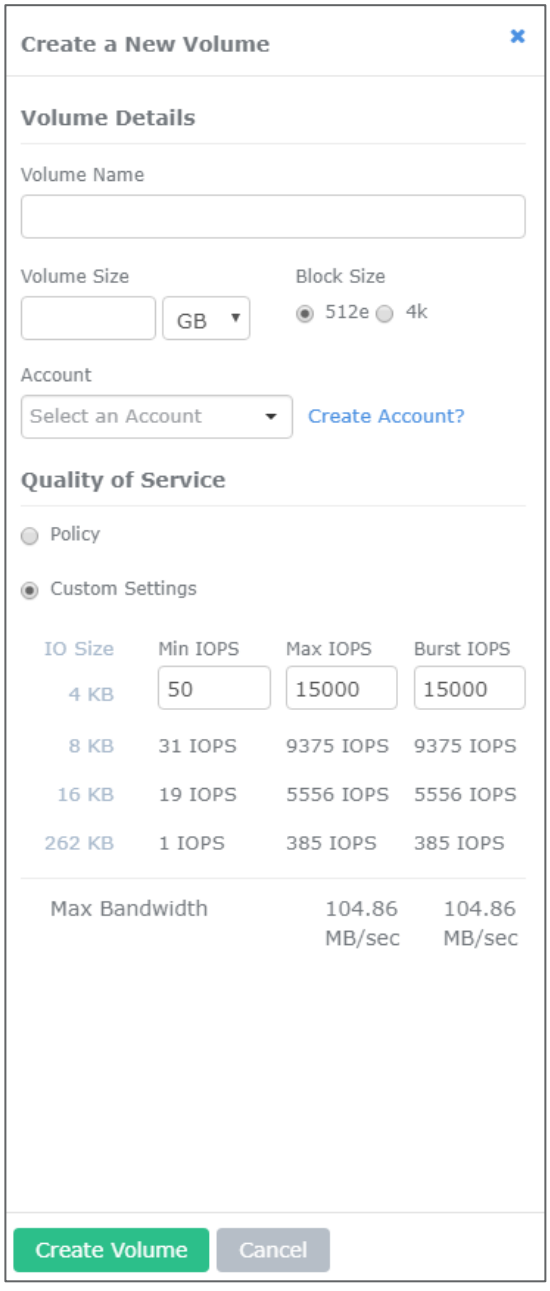
1. Go to Management > Volumes.
2. Click Create Volume.
3. Perform the standard volume-creation steps.

4. Set Quality of Service options:

- Under Policy**,** you can select an existing QoS policy, if available.
- Under Custom Settings, you can enter values or accept the default IOPS values.

Note: Volumes that have a Max IOPS or Burst IOPS value greater than 20,000 IOPS might require high queue depth or multiple sessions to achieve this level of IOPS on a single volume.

Figure 4) Custom QoS settings for a new volume.



Configuring QoS Policies

Viewing QoS Policy Details

On the Management > QoS Policies page, you can view the following information.

Table 10) QoS policy information.

Heading	Description
ID	The system-generated ID for the QoS policy.
Name	The user-defined name for the QoS policy.
Min IOPS	The minimum number of IOPS guaranteed for the volume.
Max IOPS	The maximum number of IOPS allowed for the volume.
Burst IOPS	The maximum number of IOPS allowed over a short period for the volume. The default is 15,000.
Volume	The number of volumes using the policy. This number links to a table of volumes that have the policy applied.

Creating a QoS Policy

You can create QoS policies and apply them to volumes.

1. Go to Management > QoS Policies.
2. Click Create QoS Policy.
3. Enter the policy name (up to 64 characters).
4. Enter the Min IOPS, Max IOPS, and Burst IOPS values.
5. Click Create QoS Policy.

Figure 5) Creating a new QoS policy.

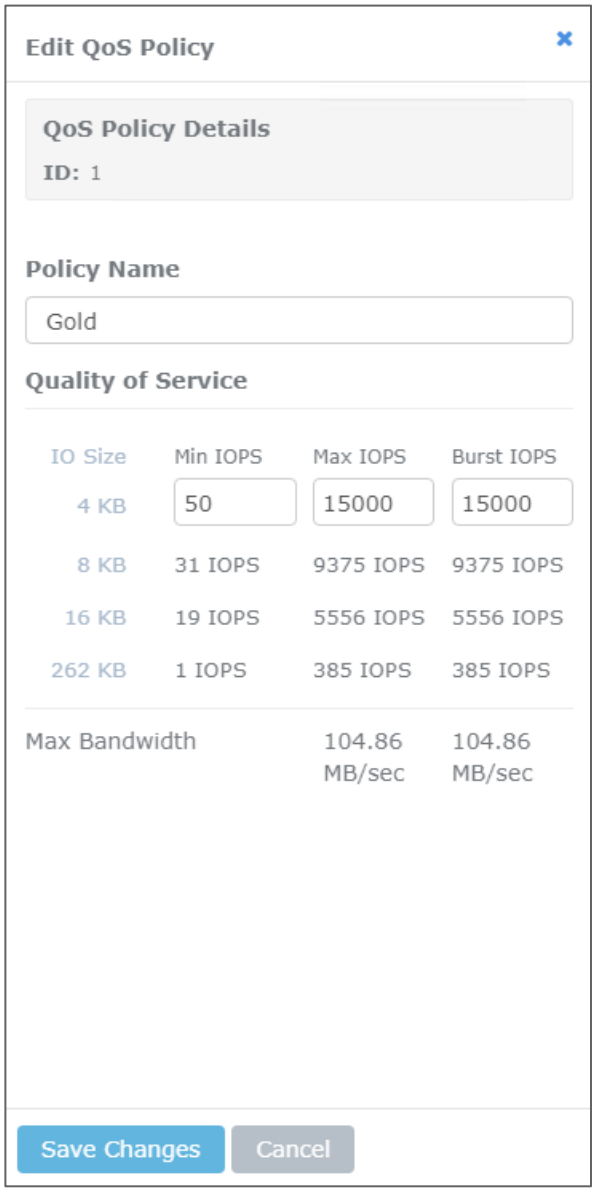
Editing a QoS Policy

You can change the name of an existing QoS policy or edit the values associated with the policy.

1. Go to Management > QoS Policies.
2. Click the Actions button (the gear icon) for the QoS policy you want to edit.
3. In the resulting menu, select Edit (the pencil icon).
4. In the Edit QoS Policy dialog box, modify the following properties as required:
    - Policy Name
    - Min IOPS
    - Max IOPS
    - Burst IOPS

5. Click Save Changes.

Figure 6) Editing a QoS policy.

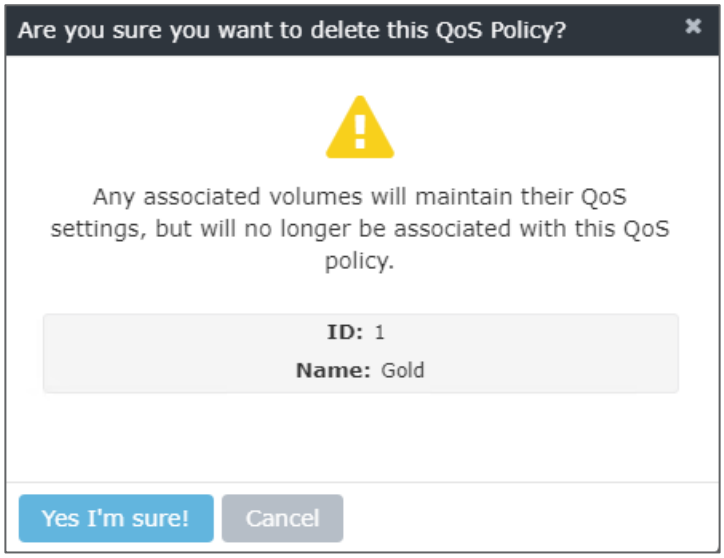


Deleting a QoS Policy

You can delete a QoS policy when it is no longer needed. When you delete a QoS policy, all volumes associated with the policy maintain the QoS settings, but become unassociated with a policy.

1. Go to Management > QoS Policies.
2. Click the Actions button (the gear icon) for the QoS policy you want to delete.
3. In the resulting menu, select Delete.
4. A message appears with a list of any volumes associated with the policy. Confirm the action.

Figure 7) Deleting a QoS policy.



5 Summary

The NetApp SolidFire storage architecture is designed to control performance independent of capacity and deliver that performance predictably to thousands of applications in a single storage infrastructure. Each volume is configured with minimum, maximum, and Burst IOPS values that are strictly enforced within the system. SolidFire QoS enables SLAs for exact performance metrics and complete control over the customer's experience. For customers, the hard QoS described in this document delivers clear expectations for storage performance and the ability to confidently deploy all applications in the cloud or on premises.

Where to Find Additional Information

To learn more about the information described in this document, see to the following references:

- SolidFire all-flash array website
  http://www.netapp.com/us/products/storage-systems/all-flash-array/solidfire-web-scale.aspx
- NetApp SolidFire QoS overview video
  https://www.youtube.com/watch?v=jiL30L5h2ik
- NetApp SolidFire Element® OS documents
  https://mysupport.netapp.com/documentation/productlibrary/index.html?productID=62480

Version History

Version	Date	Document Version History
Version 1.0	October 2017	Initial Release

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.

Copyright Information

Trademark Information

TR-4644-1017