DELL EMC VMAX3 AND VMAX ALL FLASH QUALITY OF SERVICE CONTROLS FOR MULTI-TENANT ENVIRONMENTS

ABSTRACT

This white paper describes the Quality of Service features that are available for the Dell EMC® VMAX3™ and VMAX All Flash Family storage systems. These features are available when running HYPERMAX OS 5977 Q3 2016SR and Solutions Enabler V8.3. Throughout this document, the term VMAX applies to VMAX 100K, VMAX 200K, VMAX 400K, as well as 250F, 450F and 850F storage systems.

September, 2016
The information in this publication is provided "as is." EMC Corporation makes no representations or warranties of any kind with respect to the information in this publication, and specifically disclaims implied warranties of merchantability or fitness for a particular purpose.

Use, copying, and distribution of any EMC software described in this publication requires an applicable software license.

EMC², EMC, the EMC logo, FAST, Unisphere, VMAX, VMAX All Flash and VMAX3 are registered trademarks or trademarks of EMC Corporation in the United States and other countries. All other trademarks used herein are the property of their respective owners. © Copyright 2016 EMC Corporation. All rights reserved. Published in the USA. 09/16, technical note, H15498

EMC believes the information in this document is accurate as of its publication date. The information is subject to change without notice.

Dell, EMC and other trademarks are trademarks of Dell Inc. or its subsidiaries.
# TABLE OF CONTENTS

**EXECUTIVE SUMMARY** ........................................................................................................... 4  
AUDIENCE........................................................................................................................................ 4

**INTRODUCTION** ........................................................................................................................ 5

**HOST I/O LIMITS** ....................................................................................................................... 6  
VMAX Host I/O Limits Architecture Overview .................................................................................... 6  
Bursting with Host I/O Limits ........................................................................................................... 6  
Dynamic Distribution ......................................................................................................................... 7  
Dynamic distribution modes ............................................................................................................. 7  
Alerting on when Host I/O Limits are breached ................................................................................ 8  
Host I/O Limits in Action .................................................................................................................. 9  
Creating Cache Partitions ................................................................................................................ 10  
Monitoring Cache Partitions ............................................................................................................ 12

**COPY QUALITY OF SERVICE** ................................................................................................... 13  
Replication Copy QoS ....................................................................................................................... 13

**SRDF DIRECTOR WORKLOAD DISTRIBUTION** .................................................................. 15

**CONCLUSION** ........................................................................................................................... 15

**REFERENCES** ............................................................................................................................ 16
EXECUTIVE SUMMARY

Organizations around the globe need IT infrastructures that can deliver instant access, at all times, to the massively increasing volumes of data associated with traditional transaction processing and big data use cases such as data warehousing, OLAP etc. At the same time, these organizations are trying to understand how the new generation of applications, built around the world of social, mobile, cloud, and big data (known as 3rd Platform) can be leveraged to provide economies and new business opportunities. Dell Technologies is at the heart of this discussion and is redefining Data Center Cloud Platforms to build the bridge between these two worlds to modernize and deliver the next generation Hybrid Cloud.

The best cloud based storage architectures require:

- Reduced costs by converging infrastructure and consolidating workloads.
- Massive capacity scaling.
- Massive performance scaling.
- Flexibility to handle highly fluctuating workloads yet maintain consistent service levels, all the time.
- A data protection infrastructure that can scale with a cloud scale deployment.
- A usage model that is automated and almost totally hands off.

This paper outlines the Quality of Service features of VMAX All FLASH and VMAX3 that ensure these capabilities can be delivered.

AUDIENCE

This White Paper is intended for Dell EMC customers and Dell EMC personnel.
Dell EMC VMAX3 and VMAX All Flash are incredibly well positioned to solve the CIO challenge of embracing a modernized flash-centric data center and hybrid cloud while simultaneously trying to simplify, automate and consolidate IT operations. The new VMAX isn’t just bigger, better and faster – which it is – The new VMAX architecture was designed as a data services platform that specifically addresses the new requirements of the modern data center while continuing to deliver the reliability and availability our customers have relied on for years.

With VMAX, the industry’s leading tier 1 array has evolved into a thin hardware platform with complete set of rich software data services servicing internal and now external block storage. VMAX data services are delivered by a highly resilient, agile hardware platform that offers global cache, cache partitioning (processing) flexibility, performance and High Availability at scale able to meet the most the performance needs of the world’s mission critical applications.

Dell EMC VMAX arrays continue the legacy of all Symmetrix / DMX and VMAX arrays that have come before it. It allows IT departments to automate, modernize and converge their data center infrastructure while delivering mission-critical storage with the scale, performance, availability, security and agility they’ve come to rely on for years.

Dell EMC VMAX arrays are the ideal platform for consolidation of many disparate workloads onto a single system that will provide the reliability, availability and serviceability at greater than 6x9s of reliability (With SRDF) at a scale unmatched in the industry, up to 4PBu in 4 floor tiles. To ensure consolidation and multitenancy environments have consistent performance, VMAX arrays provide a wide range of quality of service controls that enable customers to consolidate islands of block and file storage, simplify management, and reduce deployment costs by up to 33%.

**VMAX AND THE NEED FOR QUALITY OF SERVICE**

The need for consistent predictable performance is a must for enterprise customers. The architectural changes in the latest VMAX models, have ensured that this is easier to achieve than ever before with new provisioning models by Service Level in the VMAX3 and the introduction of VMAX All Flash. Service Level Provisioning ensures that storage is provisioned with a set expectation for performance, in terms of response time, and the array works to ensure that these needs are met. In addition to the changes to how storage is provisioned and consumed, key architectural changes, namely the increased processing power, ensure that there are more shared resources in VMAX arrays to serve host IO.

Architectural improvements in the New VMAX, ensures that noisy neighbors are less of a concern. CPU cores are no longer dedicated but pooled within the VMAX directors as shown in Figure 1. This pooling enables the full processing power of a VMAX directors front end capabilities to be served up to hosts via a single port ensuring that the port is not a limiting factor for performance. This is especially useful when sharing resources between different applications on the same set of front end ports.

**Note:** Additional ports may be required to scale for bandwidth needs of an application depending on port speeds.

![Figure 1. CPU Pooling in VMAX Directors](image)
bandwidth that can be consumed by an application. The VMAX has a number of Quality of Service (QoS) features that help storage administrators address these use cases.

**HOST I/O LIMITS**

Service Level Provisioning in the VMAX3 hybrid models and VMAX All Flash ensures that storage is provisioned with a set expectation for performance, in terms of response time and the array works to ensure that these needs are met. However in service providers and multitenant environments where a pay per use model is in place storage administrators may want to limit the amount of IO (IOPS) or bandwidth (MB/s) a particular application can drive on the array. To address these specific use cases, VMAX has a feature called Host I/O Limits. The HYPERMAX OS array Host I/O Limits feature allows you to define limits to enforce service levels and make application performance more predictable. The Host I/O Limits features allows the VMAX administrator to limit front-end (FE) port performance by either IOPS, Host MB per host, or both of combination of both on a Storage Group. Once set by the administrator the bandwidth and I/Os controls are then monitored by the HYPERMAX OS to ensure that they do not exceed the specified maximum bandwidth or maximum IOPs.

**VMAX HOST I/O LIMITS ARCHITECTURE OVERVIEW**

The provisioning model in VMAX utilizes the concept of a Storage Group (SG) contained within a Storage Group. Storage Groups are a collection of VMAX devices (these usually translate to individual applications storage devices. A Storage Group must exist in order to provision devices to a host. VMAX supports up to 16K Storage Groups per array). The Storage Group concept in VMAX is useful as we can reuse it for simplifying the use of VMAX data services such as local and remote replication, and also Quality of Service as described in the upcoming sections.

Host I/O Limits can be added, removed, or modified at the Storage Group level. The Host I/O Limit for a cascaded Storage Group can be added for both the parent and the child Storage Group. If a parent Storage Group has a control set, the setting is shared among all its child Storage Groups when a provisioning view is created using the parent Storage Group.

![Figure 2. Cascaded Host I/O Limits](image)

If a Masking View is created on a parent Storage Group, and both the parent and child Storage Groups have Host I/O limits, devices in the child Storage Groups will be limited by both the child and the parent Host I/O limits. In a cascaded scenario as shown in Figure 2 it is possible to oversubscribe the Host I/O limits, for example the combined limits of all the child groups can be more than the parent level. One thing to be aware of, is that you cannot set the limit of any one child group to be greater than the parent. Using the example in Figure 2 the storage admin could set any or all the child groups to the maximum of 15000 IOPs. Any attempt to set to a higher value will fail.

**BURSTING WITH HOST I/O LIMITS**

During an I/O burst the Host I/O limits feature can allow up to 1.5 of Limit for example if there is a Host I/O limit on a SG set to 1000 IOPS. The maximum credit that can be accumulated is up to 50% of the limit (500 IOPS). If no I/O was received within the previous 1 second maximum I/O that could be sent from the bucket to the VMAX would be up to 1500 IOPS in the next second. If 2000 IOPS were
received the bucket would limit the throughput to 1500 IOPS in the first second. This would exhaust credits built up and subsequent I/O would be throttled to 1000 IOPS to comply with value set.

**DYNAMIC DISTRIBUTION**

The dynamic distribution of Host I/O limits can be set across configured director ports in a port group. This allows for the Host I/O Limit allocation on each individual port to adjust to fluctuating demand. In the absence of dynamic distribution, each front end director in the port group is allocated a fixed value of Host I/O limits and these limits remain the same regardless of whether other front end directors in the port group are online or offline. For example, if the host I/O limit is 10,000 IOPS and there are two front end ports in the port group, each director will be assigned a static 5,000 Host I/O Limit.

**DYNAMIC DISTRIBUTION MODES**

Host I/O limits are shared between the front end directors to which Storage Groups are mapped and the redistribution of Host I/O limits is provided based on customer requirements. Some customers prefer redistribution of Host I/O limits among director ports and some prefer that if a director has failed, the unused Host I/O limit to be failed over to the remaining online directors.

Accordingly, three settings are available – “Always”, “OnFailure”, and “Never”.

“Always” enables the Host I/O limit to be distributed evenly in all kinds of situations, such as when a front end director fails, a director is added, or a director is removed. “Always” allows for the configured Host I/O limits to be dynamically distributed across the configured ports.

The setting “OnFailure” enables the Host I/O limit to failover when one of the front end director ports in a port group fails. “OnFailure” causes the fraction of Host I/O limits to be adjusted among the remaining online ports in if any of the ports in the port group go offline. “OnFailure” is a subset of “Always”.

The default setting is “Never” which means that dynamic distribution is disabled and no dynamic distribution setting is available for the Storage Group. Figure 3 shows setting host I/O Limits when provisioning storage with the Storage Group provisioning wizard in Unisphere, note that host I/O Limits can also be set with the REST API for Unisphere and also with Solutions Enabler Command Line Interface. Host I/O limits are set by clicking the Set Host I/O limits once you have specified the storage requirements set the limit on IOPS/MB or both and set the Dynamic Distribution policy.

![Figure 3. Setting Host I/O Limits with Unisphere Provisioning wizard](image-url)
ALERTING ON WHEN HOST I/O LIMITS ARE BREACHED

Having I/O Limits in place allows for predictable controlled levels of performance however it may be a requirement to know if applications or Storage Groups are breaching their assigned limit, it may be that they have been set too conservative or the demands of an application have changed. Unisphere for VMAX has very comprehensive alerting capabilities enabling administrators to monitor metrics that make sense to their own business needs. Figure 4 shows the alert settings administration screen, the performance threshold and alerts section highlighted in red is where we can access the alert settings for the host I/O Limits. Figure 5 shows the alerting thresholds that can be set. There are two thresholds, the first one will issue a Warning Level alert and the Second Threshold will give you a critical alert. Alerts notifications can be received via SNMP, Email or send to a syslog server e.g. Splunk.

The threshold values are set in percentage of the time exceeded during the measurement interval that a Storage Group attempted to exceed its set limit. This value interpreted as the fraction of seconds per second that the limit was exceeded – in the example Figure 5 below a warning will be sent if any Storage Group with Host I/O Limits set exceed the limit for more than 25% of the time, and if they exceed 60% a critical alert is received.
HOST I/O LIMITS IN ACTION

Figure 6 shows a simple test showing I/O Limits being enforced on a VMAX Array. IO was generated against a Storage Group with no limits in place. After some time, a limit of 50,000 IOPS was set on the Storage Group. Once the limit was set there is a drop in IO as expected as the array is forcefully throttling the host. When the limit is removed IO once again returns to the previous level. Figure 7 shows an alert generated when the host I/O Limit is being enforced for a period of time for the Oracle Storage Group.

Figure 6. Host I/O Limits in action.

![Graph showing I/O Limits](image)

Figure 7. Alerting when Host I/O Limit thresholds are breached

CACHE PARTITIONING

VMAX, has been designed to leverage a large front end cache configuration. Historically this was to maximize performance due to slow back spinning drives. Solid State Drives’s (Flash Drives) are an order 20 X faster than HDD’s but are still relatively slow compared to DRAM which is used in cache. So the DRAM architecture that has served us well, also serves us well in the All Flash platform and still provides a significant performance boost to applications with added benefits of ensuring longer life of Flash Drives. VMAX850F/X arrays can support up to 16TB of Cache, providing a huge amount of cache for the most demanding of workloads and allowing for consolidation on a scale not achievable by any other array in the industry.

For more detail on the optimization that the VMAX cache provides for flash drives please refer to the whitepaper The VMAX All Flash Storage Family – A Detailed Overview

Cache Partitioning is a feature of HYPERMAX OS which enables the storage administrator to divide the VMAXs large cache memory into multiple distinct partitions with unique names to which an applications storage devices can be dynamically assigned. Partition areas can be made static or dynamic in size. Some applications benefit more from cache than others and some applications are more cache friendly than others (i.e they have a workload that will reuse cache in an efficient manner without negatively impacting others), the Cache partitioning feature provides flexibility to the amount of floating memory that can be allocated with a high and low watermark percentage values. It also allows memory resources to be temporarily donated to other partitions when needed. The symqos command
allows creation of partitions for different device groupings in addition to the default partition that all devices belong to initially. Each partition has a target cache percentage as well as a minimum and maximum percentage. Unused cache can be donated to other partitions after a specified donation time. HYPERMAX OS 5977 Q32016 SR supports up to 16 cache partitions.

Common use-cases where cache partitioning can be useful

- Multi-Tenancy environments where there is a requirement for specific service level.
- ProtectPoint – Dynamic Cache Partitions can be leveraged on ProtectPoint level 0 (full backup) operations to limit the cache usage of the backup devices during larger copy operations.
- Limiting the usage of cache on non-cache friendly applications.
- Mainframe environments running on mixed CKD/FBA systems may wish to ring-fence the cache used by mainframe applications to ensure that cache is always dedicated for the most critical applications in instances where non cache friendly applications have been observed to consume excessive amounts of cache memory.
- Dedicating Cache for operations to be able to use more cache memory avoiding/reducing the amount of read/write misses.
- An RDF group(s) can be put in a separate cache partition to speed up remote operations.

As you can imagine Cache Partitioning provides a great degree of control, but cache partitioning needs to be used with Caution! A user of cache partitioning should have a full awareness of the entire array configuration before creating a cache partition. For e.g. total number of devices configured, total amount of cache on the box. It is recommended that users consult with their local VMAX Speed Guru if they are unsure about whether or not configuring Cache partitions will address their specific use case. In the majority of systems cache partitioning is not required as the advanced cache management algorithms of HYPERMAX OS will be able provide sufficient cache resources providing the system load is within best practice limits. The VMAX sizing process and use of the admissibility checks when adding new workloads will help ensure that this is always the case.

CREATING CACHE PARTITIONS

The cache partitioning feature can only be actively managed with the command line using the solutions enabler symqos command set. Once cache partitions are configured they can be monitored via the Unisphere for VMAX system Dashboard. Before creating a By default, only one cache partition is available on the array and this is the name DEFAULT_PARTITION. Enabling cache partitioning doesn’t make any changes to the array just yet, enabling will simply allow you to begin configuration which will then modify the array configuration and change how cache is consumed.

Creating a cache partition is accomplished via Solutions Enabler:

```
$ symqos -Cache Partitioning -name unfriendly -sid 008 create -target <TargetPercent> -min 0 -max
```

**Figure 8. Creating a Cache Partition**

* `-target` - Specifies the target cache percentage for a cache partition.
* `-TargetPercent` - The target cache percentage value for a cache partition. The target cache percentage should be more than or equal to the minimum cache percentage and less than or equal to the maximum cache percentage.

To add all devices in an applications Storage Group to an already created cache partition this is done by converting the Storage Group to a device group and adding all at once shown below in Figure 9 for Storage Group unfriendly.

```
$ symsg -sid 008 sg2dg unfriendly unfriendly
$ symqos –sid 008 -cp -name unfriendly -g unfriendly addall
```

**Figure 9. Adding a Storage Group to a Cache Partition**

The symqos analyze mode determines how much actual cache your applications require. While in analyze mode, applications will not fail due to cache partition overflow. Cache partitions cannot be changed from enabled to analyze; the cache partition must be disabled to use the analyze mode.
To run in analyze mode, set the target percent, and then set all partitions with a minimum of 0%, a maximum of 100%, and a donation time of 0 seconds in the example shown in Figure 10 there is a relatively cache unfriendly workload running against the devices in the Storage Group unfriendly, these devices have been associated with the cache partition UNFRIENDLY. The output shows that the workload is using 60% of the cache partitions target utilization which is 30% of the total system cache.

$ symqos -sid 008 -cp analyze
$ symqos  -sid 008 -cp list -settings -usage

Symmetrix ID: 000197000008

Cache Partitioning         : Analysis
Number of Partitions       : 2
Max Num of Partitions      : 16
Min Allowed Target %       : 5
Max Allowed Target %       : 90
XRC Partition State        : Disabled
XRC Partition Name         : N/A
Empty Partition State      : Remove
Time of Last Modification  : 01:53:29 PM on Mon Aug 08,2016

Cache Slots

<table>
<thead>
<tr>
<th>Partition Name</th>
<th>Min Tgt Max WP Time</th>
<th>Device</th>
<th>Used</th>
</tr>
</thead>
<tbody>
<tr>
<td>DEFAULT_PARTITION</td>
<td>0       70 100 75</td>
<td>0 37 330</td>
<td>1273615</td>
</tr>
<tr>
<td>UNFRIENDLY</td>
<td>0       30 100 75</td>
<td>0 8 9354</td>
<td>545292</td>
</tr>
</tbody>
</table>

Figure 10. Setting CACHE PARTITIONING Analyze Mode and Verifying Partition usage

Once the administrator is satisfied that the cache partition values that have been selected are appropriate for the application workload to be placed in the partition. Cache Partitioning can be enabled with the enable command as shown in Figure 11.
Figure 11. Enabling Cache Partitioning

MONITORING CACHE PARTITIONS

Once cache partitioning is set up, there are a number of ways to monitor cache usage including the heatmap views in performance analyser of Unisphere for VMAX.

Figure 12. VMAX Heatmap with Cache Partitions
Alerting can also be configured for Cache partitions to ensure that if a cache partition is reaching any thresholds a notification can be sent to ensure that the cause is understood and can be remedied.

![Figure 13. Key Performance Indicator alerts for Cache Partitioning.](image)

**COPY QUALITY OF SERVICE**

As well as providing quality of Service features that will benefit a multi-tenancy environment VMAX arrays also provide additional QoS controls that allow for storage administrators to adjust data transfer pace on specific devices or Storage Groups for operations such as TimeFinder copy and SRDF. In contrast to Cache Partitioning copy QoS is a throttling mechanism that limits or increases the copy pace of internal copy operations however slowing copy operations will make full copy operations take longer, however this reduces the number of write pending that could build up if a large copy job is executed when a system is heavily loaded with IO.

**REPLICATION COPY QOS**

VMAX local replication TimeFinder SnapVX provides the ability for customers to take snapshots of their production data at scale, it also provides the ability to link devices/Storage Groups to provide access into these point in time copies for backup and other purposes such as test and development. For detailed information on the inner workings and recommended practices of TimeFinder SnapVX [Dell EMC HYPERMAX OS TimeFinder Local Replication Technical note](https://www.dell.com/support/home/us/en/79/product-support/product/vmax-snap/vmax-snap-timesaver-technical-note).

One of the key features of SnapVX is to provide space efficient access to point in time data and sharing tracks where possible for maximum space savings. SnapVX also defaults to no copy modes when providing access to copies (by linking to target devices). There are some use cases where a full copy may be required for whatever reason and SnapVX provides for this by either taking a full copy clone in legacy mode or using a full copy linked target. HYPERMAX OS provides a copy Quality of service function that can be utilized when taking these full copies to ensure that there is no adverse effect to response times seen on the production volumes while the background copy is taking place. Different copy pace (and priority) levels can be set on certain types of devices. Possible copy pace levels on the devices are 0, the highest (this is the most aggressive and will complete the fastest), to 16, the lowest (which will effectively stop the copy process). Zero is the default.

Note: For SnapVX the QoS copy pace should be set on the **Target Devices** and it is the clone copy pace that controls this, setting on the source devices will have no effect except on restore operations where the direction of copy is to the source device from the snapshot deltas in the Storage Resource Pool (SRP). Figure 14 below shows setting the replication QoS Clone for the target devices for the Oracle_Snap_TGT Storage Group.
It should be noted that the same Copy QoS can be applied for SRDF Copy operations also as shown in Figure 15. SRDF/QoS can be applied to give higher priority to certain SRDF copy operations during bulk Migrations note: It is not usual or advised to run SRDF QoS to throttle down the throughput for production volumes running in SRDF/S mode as this may have detrimental performance impact to the source volumes. Figure 15 shows how to access the replication QoS settings in Unisphere for VMAX. Figure 14 shows how setting TimeFinder QoS for full volume copies for the Oracle Storage Group, Setting Quality of Service on SRDF is done in exactly the same way, Mirror Copy only applies to older VMAX and DMX arrays.
**SRDF DIRECTOR WORKLOAD DISTRIBUTION**

All VMAX arrays support mixed SRDF workloads on the same RDF director, i.e., you can transmit SRDF Synchronous and Asynchronous I/O across the same directors and ports. This allows even a smaller VMAX configuration to replicate data to multiple VMAX arrays over varying geographic distance. On larger systems more ports dedicated to SRDF it's worth considering to segregate Synchronous and Asynchronous traffic by director.

Figure 16 shows the cache partitioning I/O Setting accessible from the RDF Directors view under the System Dashboard in Unisphere for VMAX. Through this interface, the storage administrator can make changes to the percentage of the cache partitioning time the SRDF emulation on a director will be proportioned for Synchronous, Asynchronous and Copy Traffic. Any change made to a specific SRDF director's settings takes effect immediately for all traffic on that director emulation. If the direction of replication is expected to change at any time e.g. failover with a personality swap, it is advised to set the same settings on both arrays. The default settings are shown in Figure 16.

![Figure 16. Setting SRDF Workload Distribution default settings.](image)

Before making any changes it is always advised to review performance data so that the changes being made will make the most sense. The dashboards in Unisphere will provide a lot of information about what is happening on the array. If in doubt about what changes to make if any, to make consult with your local VMAX Speed Guru.

**CONCLUSION**

This paper was intended provide an overview of the quality of service controls available in VMAX family storage. VMAX arrays are designed for ease of provisioning to enable a short lead time until first I/O on the array. While default VMAX configurations are intended to meet the widest range of customer use cases there are often more stringent requirements in Multitenancy and Service Provider environments where pay per use models may be in place. The quality of service controls detailed in this whitepaper provide storage administrators with a comprehensive set of controls where they are able to exercise more control over I/O to the VMAX array from their customers host and segregate distinct workloads where required. For more details, please refer to the Solutions Enabler Array controls and Management CLI user guide or consult your Dell EMC SPEED guru.
REFERENCES
A full list of VMAX3 and All Flash Whitepapers and Technical Notes can be obtained here.

Solutions Enabler 8.3 Documentation Set