QV2020 and QV1020 All Flash Array
Best Practices and
Configuration Guide

LEGAL NOTICE

IMPORTANT NOTICE

Contents

Preface

This guide contains the best practices and recommendations from VIOLIN Systems to ensure optimal performances of the QV2020 and QV1020 All Flash Array. The guide is designed for use by administrators who are responsible for initial system setup and storage managers, who are responsible for configuring, monitoring, and maintaining the QV2020 and QV1020 in a production environment.

Contacting VIOLIN

To obtain additional information or technical support for VIOLIN products, or to obtain an RMA number and replacement product, contact us at:

- Phone US Toll Free: +1 855-846-5465
- Global Support Numbers available at: https://www.violinsystems.com/support-services/
- E-mail: support@violinsystems.com

Introduction

This document presents the optimal configurations for integration of the QV2020 and QV1020 All Flash Array within switched and direct-attached environments. The document will evolve over time as new information becomes available. Please check with VIOLIN Support to obtain the latest updates.

Configurations described within this document represent tested and supported environments. Other configurations may work but if problems occur VIOLIN might not be able to fully troubleshoot and provide resolution.

Intended Audience

This document is intended for IT Administrators, Desktop Specialists, Partners, and Field Services.

Hardware Configuration

QV2020 and QV1020 Components

Refer to the *QV2020 and QV1020 Hardware Guide* for a complete description of each component. This is a quick reference only.

The QV2020 and QV1020 All Flash Array has the basic external view:
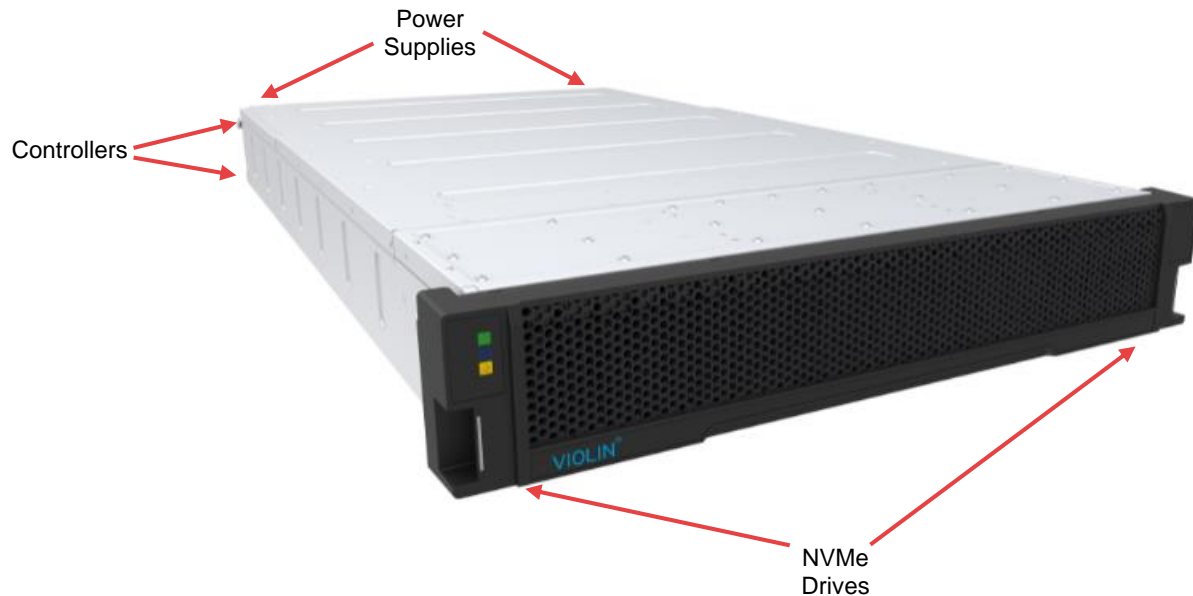
The components are (external and internal):

1. Two Power Supply Units (PSUs), providing a redundant power source
2. Dual Redundant PSUs N+1 (~2000w), Support AC 220V(180Vrms~264Vrms) input
3. Two Controllers (Fibre Channel or iSCSI) providing control and value-added features
    a. QV1020 has 2 or 4 16Gb FC ports per Controller
    b. QV2020 has 4 16Gb FC ports per Controller
    c. iSCSI 25/10G with 4 connection per Controller
4. Ten Fans
5. Supports 11 or 22 Dual-Port NVMe SSDs (solid state drives)
6. Cable management system
7. One chassis, custom-designed to house the active components of the Array
8. Bezel

QV2020 QV1020 Physical Connectivity

The Fibre Channel QV2020 has 8 16Gb FC ports, 4 per Controller, that provide FC connectivity to a fabric switch or direct-attached server. The QV1020 has 2 or 4 16Gb FC ports per Controller.

The iSCSI QV2020 and QV1020 have 8 25/10Gb network ports, 4 per Controller, with connectivity through a switch.

Any or all of the Controller ports can be cabled to provide additional paths to the initiator for higher performing IO. However, care should be taken to have a connection on each Controller to each fabric or network for redundancy.

Each LUN on an QV2020 or QV1020 has a preferred Controller, which is assigned at the time of LUN creation. Since the Controllers are active-active (any Controller can process I/O to any LUN), it is possible for a LUN to receive I/O requests from either Controller. When a LUN receives an I/O request from the non-preferred Controller, a small performance hit occurs for every I/O since the non-preferred Controller tries to forward the I/O to the preferred Controller and then sends the appropriate response back to the server. Thick and thin volumes are assigned to a Controller in an alternating manner. Dedupe LUNs are managed through one Controller to provide optimal mapping of duplicate data within the Array. If the managing Controller reboots, the other Controller will take over and become the managing Controller.

iSCSI QSFP to SFP+ fan-out Cables

Several vendors make a fan-out cable that has a single QSFP interface on one end and 4 SFP+ interfaces on   the opposite end, called fan-out cables. These work well when the QSFP is plugged into the switch and the SFP+ cables into the servers. Because each Array Controller port reports a unique MAC address, the QSFP must be plugged into the switch and the SFP+ into the Controller ports. Otherwise, what appears as a routing loop can occur since the same MAC address is visible simultaneously on multiple switch ports.

Below is an what the Mellanox QSFP to SFP+ fan-out cable looks like:

Fibre Channel Cables

For cable lengths up to 35 meters (114 feet), an OM2 cable works for all speeds from 1 Gb/sec to 16 Gb/sec. A lower bit-rate supports longer distances for the same cable. See the IBM Fibre Channel cable details link in "External References" on page 61 for exact bitrate to distance details.

Each Fibre Channel Controller has four Fibre Channel ports available for the QV2020 and two or four Fibre Channel ports for QV1020.

Fibre Channel Direct-Attached

When using one of the supported Operating System (OS) versions and supported HBAs (see the lists in VIOLIN Support Matrix), a host can connect directly to the Controller ports with a Fibre Channel cable. In a direct-attached setup, the host is required to have at least two ports per server, with each port connected to a different Controller. This ensures that the host stays online in the event of a Controller failure or a normal failover as part of an QV2020 and QV1020 firmware upgrade process.

Note: Fixed speed on the HBA and Array ports is strongly recommended for direct-attached configurations. Qlogic HBAs are the only vendor currently supported for direct-attached.

iSCSI Network

Switches that are used for iSCSI traffic should have as close to a 1:1 over-subscription ratio as possible. Significant performance impacts and possible loss of access to data can occur when Ethernet packets drop during times of heavy congestion.

Traffic Isolation

Violin Systems strongly recommends isolating iSCSI traffic from the normal business traffic (e-mail, Web, print jobs, etc.) because iSCSI traffic can consume a large amount of the available bandwidth. This also provides a degree of security as the data uses different switches than the general traffic.

Redundant Data Network Paths

There are four data networks available on the QV2020 and Q1020: A,B,C,D.

The iSCSI Array should be cabled such that each Controller is connected to the same data network.

For example, if data network A on Controller 1 is configured, then data network A on Controller 2 must be configured to provide redundancy to the Array.

Below is a Controller with the four data networks:.

Spanning Tree Protocol

In order to minimize the time it takes for the spanning tree protocol (STP) to converge, many vendors have implemented a rapid STP or RSTP protocol. Violin Systems recommends that RSTP be used because the convergence   time is between 1 and 10 seconds instead of up to 50 seconds with normal STP.

iSCSI Hardware and Software Initiators

Most operating systems are capable of mixing the supported software initiators from various sources   (Microsoft, Open-iSCSI, etc.) with hardware initiators. The software initiators cannot be run on top of the   hardware initiators, but they can co-exist and connect to different storage solutions or storage networks.

Data Network Cabling Recommendations

Most operating systems are network aware with the capability to service traffic from multiple network ports simultaneously. Multiple network connections to the same subnet results in most operating systems selecting one interface on that subnet and use it exclusively until that interface is no longer available. This behavior results in iSCSI traffic that multi-path software cannot properly handle and may not recover from failures that take the active interface offline. The complexities of maintaining routing tables in each operating system with multiple network connections, each of which is configured with a default gateway, is beyond the scope of this document.

The iSCSI controllers have five Ethernet ports each. The port labeled **Mgmt Net** is the   management network interface and it uses a standard 10/100/1G RJ-45 Ethernet cable.

Each Controller also has four QSFP interface ports labeled **Data A, Data B, Data C, Data D**. These data ports each deliver up to 25 Gb/s of Ethernet   traffic.

Violin Systems recommends separating the management network interface from the Data network interfaces to guarantee access to the management interfaces when the system is under heavy load. Isolating the   management functions from the data network can also provide a layer of security in shared environments.

Note: For the purpose of this guide, VLAN and a subnet are considered to be synonymous.

Violin Systems requires the following configuration rules for the iSCSI Array:

- Both **Data A** network ports must be on the same subnet, same with Data B, C, D
- **Data A, B. C, D** ports must <u>NOT </u>be on the same subnet; i.e. 4 separate VLANs
- A host must be connected to the same Data Network ports on each controller. Connecting to one **Data A** and one **Data B, for example,** is not supported.

Below is an example of setting up all Data Networks to be active through 2 Network switches and 2 VLANs per switch:

Active Data Networks A,B,C,D through 2 VLANs per switch
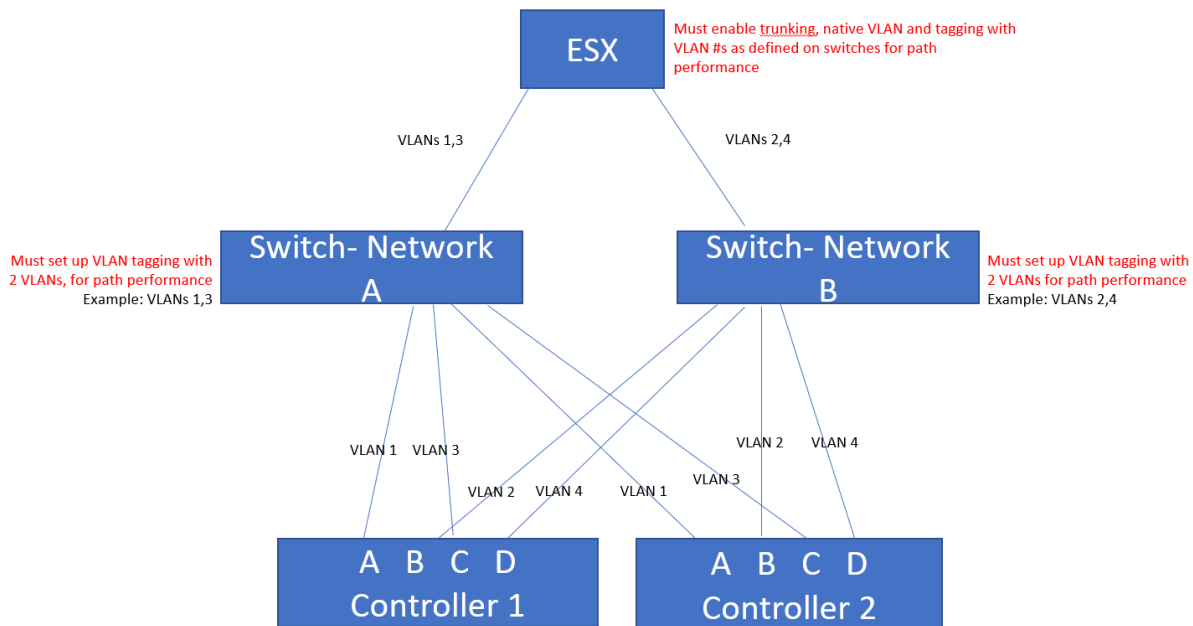
The iSCSI Array presents targets to an iSCSI initiator in the IQN format: `iqn.2004-11.com.x-io:[ArraySN]-t1`, where the text after the colon is the Array chassis serial number and a target designation. In the event that a Controller restarts or fails, load will failover over to the partner controller, where the targets are all serviced through both Ethernet ports on the surviving Controller until the partner is back in service.

Fibre Channel Switched Fabric

Fabric switches that support 8Gb or 16Gb are compatible with the QV2020 and QV1020 Array. Cisco, Qlogic, and Brocade have been qualified

Note: VIOLIN recommends the latest vendor supported switch firmware

The below configurations have been qualified by Violin Systems. Other vendor components could be compatible but issues that might occur may not be reproducible by Violin Systems to troubleshoot.

Interconnect Type	Interconnect Settings	Controller Settings
Cisco 8 & 16Gb FC Switch *Switch fw: 8.1 (1a) minimum for 16Gb*	Switch SFPs auto or fixed speed	Controller ports fixed or auto
Brocade 8 & 16Gb FC Switch *Switch fw: 7.4.2a or 2c minimum for 16Gb*	Switch SFPs fixed speed	Controller ports fixed
Qlogic 8Gb FC Switch	Switch SFPs auto or fixed speed	Controller ports auto or fixed
DAS (Direct-Attached) FC	Fibre Channel: fixed speeds on HBA with "loop preferred otherwise point-to-point" and only Qlogic HBA supported	Fibre Channel: Fixed speeds to match HBA
Cisco, Brocade, Arista, Mellanox Network Switch (iSCSI)	25G/10G, switched or Direct Attached	Data Networks must be in separate VLANs. *Reference QV2020 and QV1020 Best Practices document*

- **Note:** HBAs should be installed with the latest operating system drivers for full 16Gb support

---

General Fabric Recommendations

- Cisco switches can have either auto or fixed speeds. It is recommended the Array Controller ports match.
- Brocade switches should have switch ports set to fixed speeds, as well as on the Array Controller ports
- Brocade 8Gb switches have known issues with fill-words, causing noisy-neighbor conditions. Please refer to the above section for details.
- Switch firmware should be at the current version.
- Fabrics will negotiate faster if port speeds on the Array and the switch ports are set to fixed speeds.
- VIOLIN recommends using unique domain IDs for each switch in a solution, regardless of whether the switches are on the same fabric. This allows easier troubleshooting when viewing event logs and vendor support tools.
- 16Gb switches have introduced a feature **called RDP (Read Diagnostic Parameter**). This is disabled by default on some switches and switch versions. When this is enabled, it causes polling to the Array at a high frequency which causes unnecessary port logouts to the Array and initiator HBAs. It is recommended to DISABLE this feature on all 16Gb switch ports connected to the QV2020. In addition, several HBA vendors recommend this be disabled to the initiator ports.
- 16Gb Controller ports are recommended to have the port speeds on the Controller and the switch ports set to fixed 8Gb if the fabric switch is 8Gb or the switch ports have 8Gb SFPs.

Single Switch—Non-Redundant

It is possible to connect both Controllers and all hosts to a single fibre switch. VIOLIN does not recommend the use of a single switch as this type of configuration implies a single point of failure.

Single Switch—Redundant Fabrics

The default configuration is to connect one Controller to one fabric, and the other Controller connects to the other fabric.

The best practice is to connect at least one port from each Controller to each fabric. This provides at least one path to the preferred Controller for any given LUN on each fabric.

Multi-switch Fabrics

If there are two switches in a given fabric, cable one port per Controller to each switch and distribute the host ports across both switches in each fabric equally. Traffic between devices on the same fabric but on different switches must share the interswitch link (ISL). VIOLIN recommends having a minimum of two ISLs between each switch. If the switch supports high-speed ISLs, it is preferable to use them.

VIOLIN recommends using unique domain IDs for each switch in a solution, regardless of whether the switches are on the same fabric. This allows easier troubleshooting when viewing event logs and vendor support tools.

Very Large Fabrics

Daisy-chaining multiple switches together in large fabrics is not advArrayd as the traffic between devices on different switches must share the link and daisy-chaining reduces throughput. To alleviate this condition, use a

---

mesh topology; however, this is generally considered cost-prohibitive at any scale beyond three or four switches. VIOLIN recommends using a Core / Edge model, which has a director class switch in the center of each fabric, with top-of-rack edge switches forming the edge.

VIOLIN recommends connecting the Controllers directly to the director class switches. This minimizes the number of ISL links data has to cross to reach a given host. In the model below, a director class switch connects to multiple edge switches, which then connect to hosts. In this large fabric, VIOLIN recommends connecting each edge switch to at least two different blades on the director and connecting two ports on each Fibre Channel Controller to different blades on the director as well. This reduces the impact of a blade failure and minimizes the number of hops data has to travel through a shared ISL.

If the servers that communicate with an Array are in the same rack, it is advised to connect the Controller ports and the server ports to the same top-of-rack switches. This avoids any extra traffic on the core.

Brocade Switch Specific Recommendations

Node WWN Checking Configuration

Node WWN checking, a feature of Brocade switches, can cause some unintended paths to storage LUNs to appear on certain Controller ports. This happens because the Array uses the port WWN of the first port on Controller A as the Node WWN for all ports. See the table below as an example:

Node WWN	Port WWN	Controller and Port Number
2000001F93103688	2000001F93103688	Controller A Port 1
	2000001F93103689	Controller A Port 2
	2000001F9310368A	Controller A Port 3
	2000001F9310368B	Controller A Port 4
	2000001F9310368C	Controller B Port 5
	2000001F9310368D	Controller B Port 6
	2000001F9310368E	Controller B Port 7
	2000001F9310368F	Controller B Port 8

Consider the solution below, in which there are two ports from a server as well as the first port on each Controller connected to a Brocade switch. The zoning is configured to allow one port on the server to see Controller A and the other port to see Controller B. This should result in a total of two paths to any LUN presented from the Array to the host. When Brocade's Node WWN checking feature is enabled, the real number of paths found for each LUN is three.

Sample Switched Fabric—Two Paths Expected
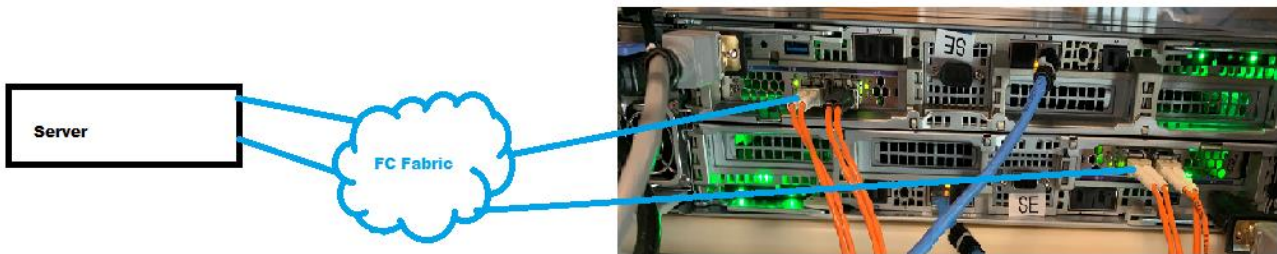
Assuming that the host WWNs are 21000024ff26415e and 21000024ff26415f, the zones would look similar to the following.

Zone Name	Member Port WWNs
MrcAPort1_HostPort1	21000024ff26415e; 2000001F93103688
MrcBPort5_HostPort2	21000024ff26415f; 2000001F9310368C

Brocade switches can build zones based on either Port WWNs or Node WWNs. When a zone is created that contains a member with identical Node and Port WWNs, the zone is assumed to be using Node WWN zoning; therefore, all other ports with the same Node WWN are added to this zone.

The Node WWN checking option impacts any zones containing the first port of the first Controller on any Array. This can be avoided by either using a different port on the first Controller of the Array or by disabling the Node WWN checking on each Brocade switch in a given fabric.

The recommended best practice is to set **zoning.check.nodeNameDisabled** to 1, which disables checking the node WWN.

Each Fibre Channel device has a minimum of two worldwide names (WWNs). One is assigned to the device and is the node WWN, and the other is assigned to the Fibre Channel port. A two-port Fibre Channel HBA has a common node WWN for the HBA, and each port has a unique port WWN. The Array WWN assignment mechanism uses the same value for both the node WWN and the WWN of the first port on the first Controller.

Consider a Brocade fabric with one switch and Node WWN checking enabled (the default behavior) that has an Array connected and zoning set up per the following list:

- HBA Port 1 to M
- RC A Port 1 (Port WWN)
- HBA Port 2 to Controller B Port 5 (Port WWN)

The expected number of paths to a LUN should be two; however, the observed number of paths is three. HBA= Port 1 sees both Controller A Port 1 and Controller B Port 5, because the zoning is built by node WWN first and then by port WWNs. Controller B Port 5 has the same node WWN as Controller A Port 1's port WWN and so is included in the zone.

Use the following procedure to disable and reconfigure the switch parameter:

WARNING: *This procedure disrupts I/O through the switch until the configuration process is complete.*

---

1. Disable the switch by typing the following line at the switch command prompt: switchdisable
2. Configure the setting using the following steps.
3. At the switch command prompt, type:

   `Configure`

4. Select Zoning Operation Parameters from the menu.
5. Change the Disable NodeName Zone Checking to one.
6. Re-enable the switch by typing the following line at the switch command prompt:

   Switchenable

7. Check the setting by typing the following line at the switch command prompt:

   configshow | grep zoning

8. Verify that the switch displays **zoning.check.nodeNameDisabled:1**.

Note: If the previous configuration utilized zones, review the configuration to ensure that the change correctly represents your zoning preferences.

Brocade Switch Fill-Word Settings

Fibre Channel switches maintain synchronization between connected devices using fill-words. These fill-words are Fibre Channel frames that have no customer data in them but contain special bit patterns that indicate that the frame is a filler to maintain synchronization when the port is not actively in use. On the 1, 2, and 4 Gb/s fabrics, the fill word frame was named "**IDLE**," and so the fill-word mechanism was referred to as "**idleidle**." Other fill words could be named "**ARB(F0)**" or "**ARB(FF)**." At the higher bitrates required to provide 8Gb/s fabrics, the pattern of bits for the "**idle-idle**" frame could potentially impact adjacent ports. As a result, 8Gb/s devices mostly use "**ARB(FF)**" as the fill-word but will fall back to the older "**idle-idle**" if "**ARB(FF)**" does not work.

If "**ARB(FF)**" is not possible, the switch reverts to "**idle-idle**". The Fibre Channel QV2020 and QV1020 maximum bitrate is 16 Gb/s, and it does understand the "**ARB(FF)**" fill-word, so there is nothing to change for a Fibre Channel QV2020 and QV1020 connected to an 8Gb switch.

Brocade offers a **portcfgfillword** command to change the default fill-word if needed. Refer to Brocade's documentation and technical support if it is believed that a change to the fill-word behavior is required. As shown below, mode 3 is the default for most Brocade 8Gb switches; however, some early versions of this switch shipped from Brocade with the default mode set to 0. An upgrade to the fabric OS will not change this parameter. It is recommended to change the fill-word to mode 3 for all ports unless otherwQV2020 indicated by Brocade or the hardware connected to a specific port.

```
admin> portcfgfillword

Usage:
portCfgFillWord PortNumber Mode [Passive] Mode: 0/-idle-idle (old default)
1/-arbff-arbff
2/-idle-arbff
3/-aa-then-ia (default)
Passive: 0/1
```

Fibre Channel Zoning with the QV2020 and QV1020

Think of Fibre Channel zones as virtual SCSI cables with two or more connectors (up to the maximum number of zone members supported by the switch). If a zone does not connect two devices explicitly, then the devices do not communicate unless the switch is configured for open zones. Open zoning allows all devices to connect and communicate with all other devices since the storage products (targets) are required to implement a LUN masking model, which selectively presents storage to individual hosts. If there are older storage devices that do not allow selective presentation, then zoning is required to provide the isolation necessary to make things work as expected.

There are several Fibre Channel zoning models, each of which has its own set of challenges and benefits. The most common zoning models are open-zoning, single-host, single-initiator, and one-to-one. One of the largest impacts the zoning model has is reducing the number of fabric-wide events that happen when a change is made. For example, rebooting a single host causes each Fibre Channel port to log out of the fabric. Log back in to the fabric, and then connect to storage devices and begin I/O. Every time a port logs into or out of a fabric, a registered state change notification (RSCN) event occurs, which is sent to every device already connected to the fabric. If the zoning does not handle this to prevent the broadcasts, server performance and access to storage can be negatively impacted.

Fibre Channel Interconnect (Controller). The QV1020 has option of 2 ports

In the sample fabric below, the host port WWNs are 21000024ff26415e and 21000024ff26415f, and the Array port WWNs are 2000001F93103688 and 2000001F9310368C. It is assumed that under each zoning model described, each host port sees both Controller ports for redundancy.
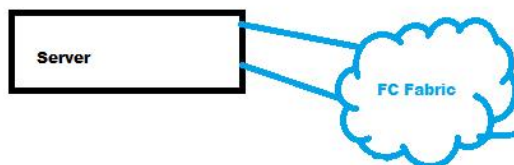
Open Zoning

Open zoning is the simplest and easiest Fibre Channel fabric model to use, because under this model all nodes connected to the fabric automatically see each other. Open zoning increases the number of devices that a host port must scan during boot or bus rescan operations. This can increase the host boot times, especially when using boot from SAN. In contrast, this model requires very little documentation to manage, and zone configurations do not need to be backed up. This model is most appropriate for fabrics consisting of a single Fibre Channel switch due to the risk of RSCN broadcasts taking hosts offline.

Host-Based Zoning

Host-based zoning is the next simplest to implement. In this model, all target ports that a given host needs to access are included in the same zone with all initiators from the same host. This model reduces the RSCN broadcasts to only the zones containing the device that logged into or out of the fabric. In situations where one host port has a problem and sends out a lot of packets incorrectly, it is possible that the healthy host port will also be impacted because the same zone includes both host ports. This model has the fewest possible number of zones to manage, back up, and document. This model is most appropriate for smaller fabrics of just a few switches.

The host-based zone membership for the fabric would look like the following table:

Zone Name	Members
HostA_Array	21000024ff26415e; 2000001F93103688; 21000024ff26415f; 2000001F9310368C

Single Initiator Zoning

Single initiator zoning would be identical to host-based zoning if the host has only one Fibre Channel port. In cases where the host has two or more ports, there is one zone for each initiator port that contains all targets with which the initiator port should be able to communicate. This limits the impact that any host port can have on any other host port and also limits the RSCN broadcasts to the zones containing the port that is logging into or out of the fabric. This model has more zones to manage the host-based zoning, which can increase the documentation and configuration back-up requirements.

The single initiator-based zone membership for the fabric would be similar to that listed in the following table:

Zone Name	Members
HostAPort1_Array	21000024ff26415e; 2000001F93103688; 2000001F9310368C
HostAPort2_Array	21000024ff26415f; 2000001F93103688; 2000001F9310368C

One-To-One Zoning

One-to-One zoning is the most controlled model. Under this model, each zone represents a single virtual SCSI cable with exactly two connectors. If a host has two initiator ports per fabric and there are three storage devices with two ports on each fabric, then there are 12 zones per fabric per server. Obviously, this can get to be a very large number of zones very quickly, but RSCN broadcasts are kept to an extremely narrow set of devices, so the impact of a failing Fibre Channel port is very small.

The One-To-One based zone membership would be similar to that listed in the following table.

Zone Name	Members
HostAPort1_MrcAPort1	21000024ff26415e; 2000001F93103688
HostAPort1_MrcBPort5	21000024ff26415e; 2000001F9310368C
HostAPort2_MrcAPort1	21000024ff26415f; 2000001F93103688
HostAPort2_MrcAPort5	21000024ff26415f; 2000001F9310368C

VIOLIN Zoning Recommendations

VIOLIN recommends the use of Single Initiator or One-To-One zoning for most applications to prevent RSCN broadcasts from negatively impacting an environment and to reduce the risk of accidentally presenting storage to the wrong server(s).

Note that this zoning covers Array to servers and does not apply to zoning requirements for clusters, storage virtualizers such as Datacore, IBM's SVC, or QV2020 or QV1020 Mirroring. These types of zones may need to have multiple initiators in the same zone. Refer to the appropriate user guide for the cluster, storage virtualizer, or mirroring tool for the specific zoning configurations needed.

Fibre Channel HBAs

A host bus adapter (HBA) is used to connect a server to a storage solution, such as an Array. An HBA can have one, two, or four ports each and can communicate at 8 or 16 Gb per second. An HBA has several configurable parameters that control how it behaves with storage solutions. Different HBA vendors may have different terms for these parameters; please refer to vendor documentation for exact details on how to set these values. The individual parameters that are commonly adjusted for an Array are described below.

Note: When storage products have conflicting HBA settings, we recommend that different HBAs be used. This may apply to storage from different vendors or to storage products from the same vendor with different requirements.

VIOLIN recommends:

- Having at least 2 FC ports in each server, which includes at least one path to each Array Controller.
- Running the latest drivers and firmware as recommended by the HBA vendor.

Qualified HBAs

The HBAs that are tested and supported include:
- Qlogic 8Gb and 16GB
- Emulex 8Gb

HBAs are recommended to have the current firmware and drivers supported by the vendor.

Connection Options (QLogic specific)

An HBA can be connected to multiple types of ports. QLogic offers "Loop Only," "Point to Point Only," and "Loop Preferred, Otherwise Point to Point." Refer to the SNIA dictionary (http://www.snia.org/education/ dictionary) for more information on arbitrated loop and point-to-point connections.

Direct-Attached

VIOLIN recommends that a QLogic HBA be set to "**Loop Preferred, Otherwise Point to Point**." This option is required for direct-attached solutions due to a known compatibility issue when using "Point to Point Only." The Array bypasses Loop and logs in with Point to Point.

QLogic defines the connection options as follows:

0—Loop Only

1—Point to Point Only

2—Loop Preferred, Otherwise Point to Point (default)

3—Point to Point, Otherwise Loop

Data Rate

An HBA port is normally set to auto-detect the maximum speed possible when connected to a switch or storage port. A Fibre Channel Array can use auto-speed negotiation.

Direct-Attached

VIOLIN recommends that the Controller and HBA ports be set to a fixed data rate when directly connected together (no switch between).

Transfer Size

Most HBA vendors have the ability to set the maximum transfer size value. QLogic's command line tools calls this value the "tsize," Emulex calls it the "LimTransferSize," and Brocade calls it "IOsize." Decreasing this value can reduce the time required to complete each I/O; however, changing it also increases queue depth since the number of requests to transmit any given block of data will be larger. If possible, it is recommended to test—not in production—various combinations of transfer sizes with the recommended max queue depths and determine what is an acceptable level of latency for the maximum queue depth available.

Adapter BIOS/Boot from SAN Settings

The Adapter BIOS is disabled by default. In order to boot from a SAN device, the adapter BIOS must be enabled and configured to boot from a specific storage device. Refer to the HBA vendor's support documentation for instructions on doing this.

Vendor-Specific Parameters

QLogic

HBA Port Settings			
QLogic HBA Setting Name	Default	Fabric Recommended/ Required	Direct-Attached Recommended/ Required
Connection Option	2 - Loop Preferred, OtherwQV2020 Point-to- Point	2 - Loop Preferred, OtherwArray Point-to-Point	2 - Loop Preferred, OtherwArray Point-to-Point
Data Rate	Auto	Auto	8 /16 (fixed preferred)
Frame Size	2048	2048	2048
Hard Loop ID	0	0	0
Loop Reset Delay (seconds)	5	5	5
Enable Host HBA BIOS	Disabled	Enable (For Boot From SAN)	Enable (For Boot From SAN)
Enable Hard Loop ID	Disabled	Disabled	Disabled
Enable Fibre Channel Tape Support	Enabled	Enabled	Enabled
Operation Mode	0 - Interrupt for Every I/ O Completion	0 - Interrupt for Every I/O Completion	0 - Interrupt for Every I/O Completion
Interrupt Delay Timer (100ms)	0	0	0
Execution Throttle	16	256* (65535 for Windows)	256* (65535 for Windows)
Login Retry Count	8	8	8
Port Down Retry Count	30	30	30
Enable LIP Full Login	Enabled	Enabled	Enabled

Link Down Timeout (seconds)	30	30	30
Enable Target Reset	Enabled	Enabled	Enabled
LUNs Per Target	8	<Desired Number of LUNs>	<Desired Number of LUNs>
Out of Order Frame Assembly	Disabled	Disabled	Disabled

*Depends on the HBA model, driver, and software being used.

Emulex

Emulex HBA Settings			
Parameter Name	Default Name	Fabric Value	Direct-Attached
Link Timeout	30	30	Not Supported
Node Timeout	30	30	Not Supported
Queue Depth	32	254	Not Supported
Queue Target / LUN	0	0	Not Supported

Cisco UCS

The adapter defaults for the HBA are shown below. The defaults apply to all HBA profiles, except VMware. Also, the default queue depth (I/O Throttle Count) is set to 16 by Cisco. While this does work, it can impose some significant performance impacts on I/O intensive applications, such as SQL Server, and virtualization environments, such as VMware. VIOLIN recommends changing this value as well.

Options

FCP Error Recovery:	◉ Disabled ○ Enabled	
Flogi Retries:	8	[0-infinite]
Flogi Timeout (ms):	4000	[1000-255000]
Plogi Retries:	8	[0-255]
Plogi Timeout (ms):	20000	[1000-255000]
Error Detect Timeout (ms):	**2000**	
Port Down Timeout (ms):	30000	[0-240000]
Port Down IO Retry:	30	[0-255]
Link Down Timeout (ms):	30000	[0-240000]
Resource Allocation Timeout (ms):	**10000**	
IO Throttle Count:	16	[1-1024]
Max LUNs Per Target:	256	[1-1024]
Interrupt Mode:	◉ Msi X ○ Msi ○ Intx	

UCS M81KR HBA Parameter Defaults

Minimums recommended by VIOLIN:

- Port Down Timeout (ms): 10000 (*VMware only*), 30000 (others)
- Qdepth: 64
- I/O Throttle Count : 256

iSCSI Attributes

CHAP

The iSCSI Array can be configured to enable CHAP security. The CHAP security model can either perform authentication of the initiator by the target (one-way) or it can authenticate the initiator by the target and then authenticate the target with the initiator (two-way or mutual authentication). The iSCSI Array can use either authentication model as long as the CHAP secret meets the criteria of being 12–16 characters long. If using mutual authentication, the CHAP secrets must be different.

CHAP can be enabled using MAESTRO. Reference the QV2020 and QV1020 All Flash Array User Guide for details.

> *Note: The iSCSI Array uses a single username and password combination for all hosts. The use of separate username and password combinations for each host is not currently supported.*

Jumbo Frames

The iSCSI Array supports "normal" Ethernet frames of 1500 bytes or "jumbo" frames of up to 9000 bytes. When properly implemented, jumbo frames can reduce the amount of overhead needed to send large blocks of data; however, the apparent performance impact of this reduction is not usually visible until traffic exceeds 50% of the bandwidth available. In order for jumbo frames to function correctly, every single device on the iSCSI network must support jumbo frames. If any device on the network, especially routers or switches, does not support jumbo frames, then the performance can get worse because the jumbo frames must be broken apart and sent as normal frames.

Reference the QV2020 and QV2021 All Flash Array User Guide for instructions on enabling Jumbo Frames thorough MAESTRO.

Note:. Violin Systems strongly recommends having iSCSI networks physically separated from other traffic when using jumbo frames.

Warning: In the event that Jumbo Frames are not properly configured on the Ethernet switches and routers between two end points, it is possible that storage will not be accessible to the host.

To test whether Jumbo Frames are properly configured on the switches, use the **ping** command with options to prevent fragmentation and a packet size larger than 1500 bytes. Examples are shown below for Linux, Windows, and VMware. Please refer to the documentation for other operating systems as command line options vary.

Linux
```
ping –M do –s 8972 [destination IP]
```

Windows
```
ping –f –L 9000 [destination IP]
```

VMware (software initiator)
```
vmkping -d -s 8972 [destination IP]
```

If Jumbo Frames are enabled, the ping command output will look normal. If the destination is not reachable with Jumbo Frames, there will be responses such as "Request timeout for icmp_seq 0" or "Request timed out." Multiple physical ports on the same subnet should have the same MTU to prevent confusion in the networking path.

Switch Vendor Notes

Warning: *Mellanox switches require the SAVE button (upper right, in picture below) to be clicked after a configuration change. Otherwise the change is lost on a switch reboot. This will result in loss of access to the Array LUNs due to mismatched interconnect speeds.*
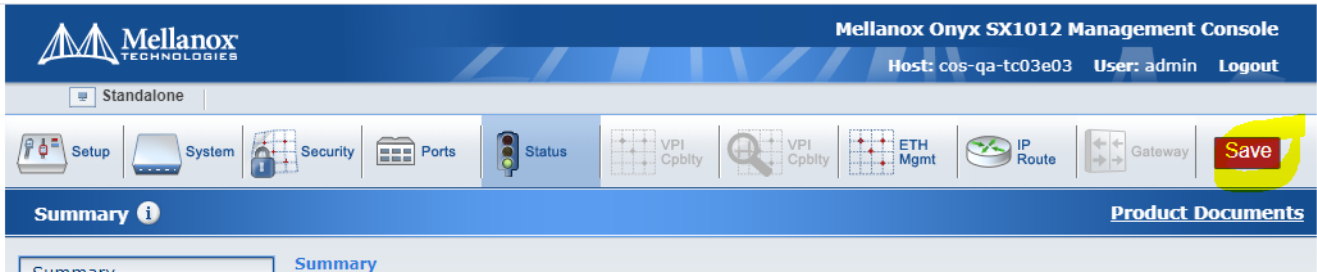
In general, reference the switch vendor documentation for details on setting MTU values for Jumbo Frames.

Array Feature Recommendations

Data Reduction (Dedupe) – Achieving Optimal Results

Volume type of dedupe is the default type. Due to the internal mapping of dedupe volumes, all IO for dedupe volumes will occur through the XO Controller, which is typically Controller 1 in a newly initialized Array. In the event of a reboot of the XO Controller, dedupe volumes will failover to the other Controller, per the standard Array failover process.

Migrating an existing VMware environment to the QV2020 or QV1020

- Use Storage vMotion to migrate the VMs to the QV2020 or QV1020
  - During the Storage vMotion wizard select a different destination virtual disk type
    - Thin for best results of dedupe and UNMAP support
    - ThickEagerZero for slightly better performance and equally good results of dedupe but no support for UNMAP
    - ThickLazyZero is not recommended as it provides no benefits and it was created for slower hard drives so they would initialize faster

New deployments

- For Windows on VMware or Linux on VMware
  - Create a Thin or ThickEagerZero virtual disk
    - Thin for best results of dedupe and UNMAP support
    - ThickEagerZero for slightly better performance and equally good results of dedupe but no support for UNMAP
    - ThickLazyZero is not recommended as it provides no benefits and it was created for slower hard drives so they would initialize faster
  - Format volume with 8k cluster size or larger
- For physical Windows or physical Linux
  - Format volume with 8k cluster size or larger

Existing volumes

(not a new deployment or migration to QV2020 or QV1020)

- For Windows on VMware or Linux on VMware
  - Create a Thin or ThickEagerZero virtual disk
    - Thin for best results of dedupe and UNMAP support
    - ThickEagerZero for slightly better performance and equally good results of dedupe but no support for UNMAP
    - ThickLazyZero is not recommended as it provides no benefits and it was created for slower hard drives so they would initialize faster
  - Format volume with 8k cluster size or larger
  - Migrate data from old volume to new volume
    - Ie) Robocopy or equivalent
- For physical Windows or physical Linux
  - Map new volume from the SAN
  - Format volume with 8k cluster size or larger
  - Migrate data from old volume to new volume such as Robocopy or equivalent

Example: Changing allocation unit size on Windows

Allocation Unit Size = 8192 (Windows example below) or larger
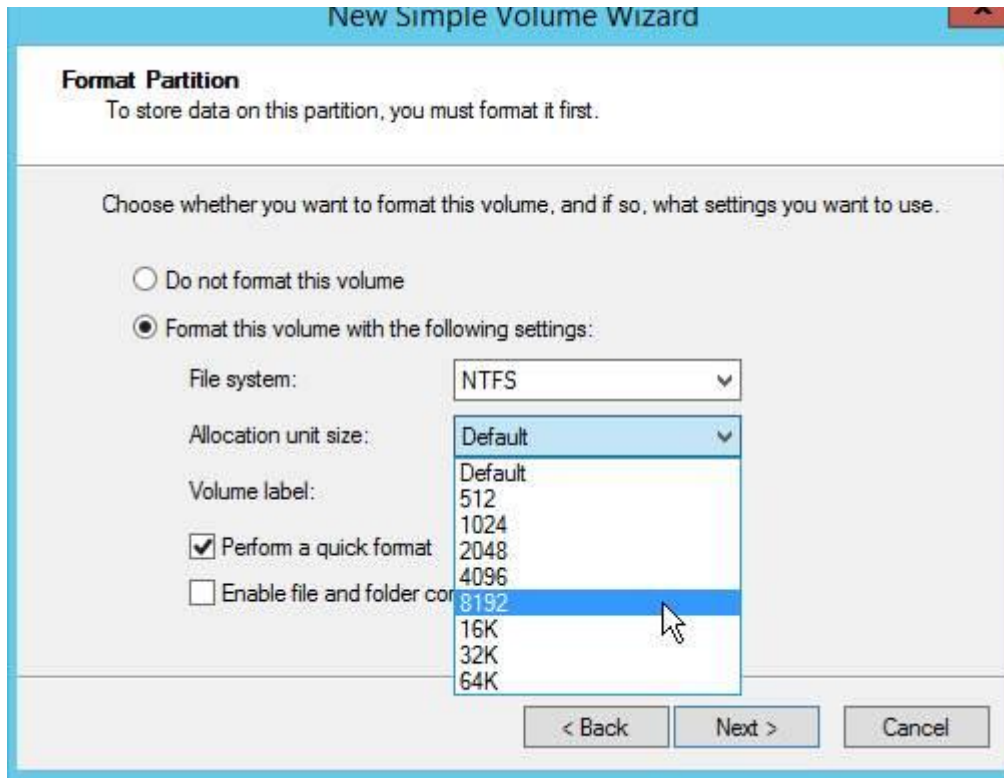
Unmap/Space Reclaim

Upon deletion of data, space can be reclaimed on Thin (or Dedupe) storage volumes by using an operating system feature often called "unmap" or "Delete Notify". The feature is supported on the QV2020 and QV1020. Once the reclaim command is issued to an Array, the capacity will be regained. For dedupe volumes, it can take time to complete, as it is a background process within the Array. In addition, if the data was duplicate data, the dedupe ratio will reflect the removal of data.

Windows 2008 does not have a reclaim feature. Space will not be reclaimed on the Array. A third-party tool is required to reclaim the space on the storage. Deleting the thin volume on the Array will reclaim the capacity and adjust the dedupe ratio accordingly.

Windows 2012 and Windows 2016 support reclaim capacity and it is automatic. This feature is called Delete Notify and it is enabled by default. Data in the Recycle Bin will not be reclaimed until the Recycle Bin is emptied.

To verify this feature is enabled:

```
fsutil behavior query disabledeletenotify
If "0" then "Delete Notify" is Enabled (UNMAP)
If "1" then "Delete Notify" is Disabled (no UNMAP)
```

To enable:
```
fsutil behavior set DisableDeleteNotify 0
```

VMFS 5 supports UNMAP but it is not automatic. Manual commands must be issued. Reference VMware documentation for details.

Example:
```
vmkfstools -y /vmfs /volumes /<datastore>
```

To check unmap progress:
```
esxtop u f o
```

VMFS 6 supports UNMAP and it is automatic. However, there is a known VMware issue that **partially** unmaps data. So, data is not completely deleted off the storage and capacity is not fully regained.

VMware 6.7 is expected to fix this bug, however current VIOLIN lab testing proves, issues still exist as of May

2018 updates. Issuing manual unmap commands may reclaim additional capacity but it is non-deterministic based on VIOLIN lab testing. Reference the VMware documentation for details.

Example:
```
esxcli storage vmfs unmap -l <datastore>
```

or
```
esxcli storage vmfs unmap   -u <uuid>
```

To check unmap progress:
```
esxtop u f o
```

Snapshot Schedules

Schedules that coincide at the same time on multiple volumes will have slight overhead in performance due to background thinning. It is recommended to stagger these intervals to avoid concurrent operations.

Reference the QV2020 and QV1020 All Flash Array User Guide for details on snapshots and schedules.

Promoted Snapshot

Snapshots are not accessible by the client/server to preserve the original data. If data needs to be read or written from a snapshot, the snapshot must be "promoted". Promote will clone the data from the snapshot and create a new volume. The new volume is independent from the original parent volume of the snapshot. Data written on the promoted volume will not be written to the original snapshot nor the parent volume of the snapshot. A snapshot can be promoted through the Volume Snapshot page, by selecting the details on the snapshot.

Defragmentation of Array Volumes

Different operating systems have different recommendations and tools regarding how to defragment the file system. VIOLIN recommends referring to the operating system vendor regarding best practices for defragmentation tools and procedures. This is especially important for virtual server environments (VMware, XenServer, Hyper-V, etc.).

RAID Impacts on Performance

QV2020 and QV1020 volumes can be created using RAID-5 or RAID-1 to protect customer data. The choice of which RAID level to use depends largely on what the customer workload is and how frequently data is written to the volume. All RAID-5 implementations operate on similar principles, so all RAID-5 volumes will have some version of a write penalty.

The RAID-5 Write Penalty

Consider a RAID-5 volume that stripes data across five drives, illustrated below. If an application writes data to some of the bytes in the block represented by B2, B2 is read into memory along with the parity data in block P. The data is updated in memory to reflect the writes from the application, and a new parity value is calculated. The changes to block B2 and P are then written to disk. One write (one I/O) from an application leads to reading from two disks and writing to two disks for a total of four I/Os. This increases if the application modifies more than one block (parts of B2, B3, and B4, for example). If the application is performing a significant amount of random I/O, the combined impact can overwhelm the back-end storage, and some applications may report that the storage has failed or even experience unacceptably high latencies.
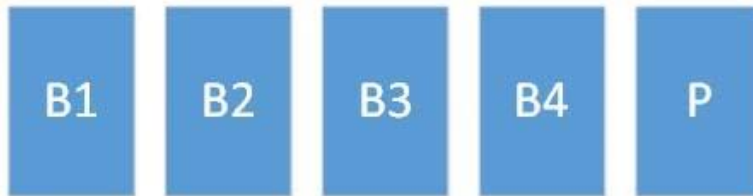
Sample 5-drive RAID set

Using the example above and assuming that each drive can serve up to 200 IOPS, a theoretical limit of 1,000 IOPS is available. Given the worst case of four IOPS per write on RAID-5, the maximum IOPS for this configuration would be 250 write IOPS on RAID-5. Any solution that regularly writes fewer than 250 IOPS will likely not encounter any significant difference between writes to RAID-1 and RAID-5 volumes. If there is a mix of read and write IOPS contributing to the load on this system, then the read IOPS plus the write IOPS (with write penalty) cannot exceed the total maximum IOPS for the system.

If the volume were RAID-1 instead of RAID-5, it would be a simple matter of two writes (one to each disk) with no need to calculate parity or write parity bits to disk. The cost, of course, is that RAID-1 occupies twice the disk space because it keeps a full mirror of all data written to disk by an application.

Due to the number of variables involved, it is nearly impossible to create a generic set of rules to determine which RAID level is best for a given environment. In general, it is safe to assume that the write performance of RAID-5 will be 2/3rd the performance of RAID-1.

Sequential writes—writing many blocks in order—can help avoid the RAID-5 write penalty because the entire stripe is modified at once. In the Sample 5-drive RAID set illustration above, sequential writes would make changes to all data on B1, B2, B3, B4 and also update the parity block. Because most storage Arrays have some form of caching capability, the cache would see the writes for each drive coming in and merge them into a single, larger write transaction.

VIOLIN recommends the following for selecting RAID types for volumes:

- For applications that make small, infrequent writes, use whichever RAID level gives the customer an acceptable level of data protection.
- Sequential I/O workloads, such as for database transaction logs, are OK to use on RAID-5 or RAID-1 volumes.
- Databases (including Temp DB) should not be on RAID-5 due to the high random I/O usage patterns.
- RAID-5 is not recommended for high amounts of small I/O writes due to the 4x back-end I/O impacts. These environments may be represented by but are not limited to VDI with VMs busily writing lots of data to the Array, with many hosts, each doing a small amount of writes.

Write Caching

Each Array volume defaults to write-back caching. The Array will automatically switch to write-through caching if hardware issues occur that impact the ability for write-back cache to save data. Write-through caching is not user configurable.

Host Limits

The QV2020 and QV1020 is limited to 320 host port WWNs logged into the Array at any one time.

*For example: if a customer environment has 25 physical servers and none are running virtual machines, each with two HBAs, then the total number of host port WWNs used is 50. If these servers each run four virtual machines with two NPIV ports presented to the Array, then the number of host port WWNs used is 250 (4 VMs * 2 ports * 25 hosts + 2 ports per server * 25 servers).*

Host Execute Throttle/Queue-Depth

The Execution Throttle specifies the maximum number of commands (I/Os) executing on a per target port (storage port) basis. When a port's Execution Throttle value is reached, no new commands (I/Os) are executed until the current command finishes executing to the specific target port. Commands can continue to be executed for other target ports that have not reached their Execution Throttle threshold.

The Queue Depth specifies the number of outstanding I/Os per LUN. In a running environment, the queue depth will not go higher than the execution throttle.

All storage solutions have a queue depth threshold value, and if the combined I/O of all connected hosts exceeds that value, the storage returns a busy response.

QLogic's HBA parameters allow the execution throttle to be set as high as 65,535, which exceeds the Array queue depth, assuming the servers and applications can generate sufficient I/O. The Emulex HBA tools allow a queue depth of 254 for Windows hosts and 128 for Linux.

As a best practice, set the execution throttle to the highest setting. Then adjust the queue depth. For most workloads, the default queue depth of 32 should be sufficient.

Operating System Considerations

Performance and upgrade options for the Array vary depending on the operating system and applications in use. Performance can be impacted by caching mechanisms in each OS, and Array firmware upgrades may require downtime under certain conditions.

Multipathing

Violin Systems recommends operating system multipathing to the Array LUNs. This will provide availability of the storage during common path loss scenarios, including firmware upgrade.

When configuring multipath to identify the Array LUNs, the below Product IDs are required:

Array Product IDs:

      ISE4400 (Fibre Channel QV2020 and QV1020)

      ISE4401 (iSCSI QV2020 and QV1020)

Array Software Upgrades

When the Array software upgrade occurs, there is a period of time while cache is being transferred between Controllers when no I/O is possible. This could take as long as 30 seconds before IO resumes, so multipath loss timers need to be adjusted to not trigger a device or path failure scenario.

Before attempting to upgrade an Array, check the multipath configuration:

1. Verify each Controller has active paths to the host running IO to the Array.
2. Verify the operating system multipath recommendations listed in this document have been applied.
3. Verify the HBA recommendations listed in the prior section have been implemented. This will ensure proper handling of path failover.

Note: Failure to have a properly configured multipath solution will likely cause disruption of I/O during firmware upgrades.

Upgrade Types

Online Upgrade

An online upgrade provides the ability to upgrade the Array software while IO continues.

Multipath and HBA parameter recommendations outlined in this document should be verified prior to executing an online upgrade. In addition, each Controller must have at least one active path to the hosts running IO/applications.

Offline Upgrade

An offline upgrade is when the Array upgrade is performed while the connected hosts are shut down or the applications that use the storage are stopped. This method is required if applications, file system, or operating system cannot survive the required path timeout of 30 seconds or less.

VMware ESXi

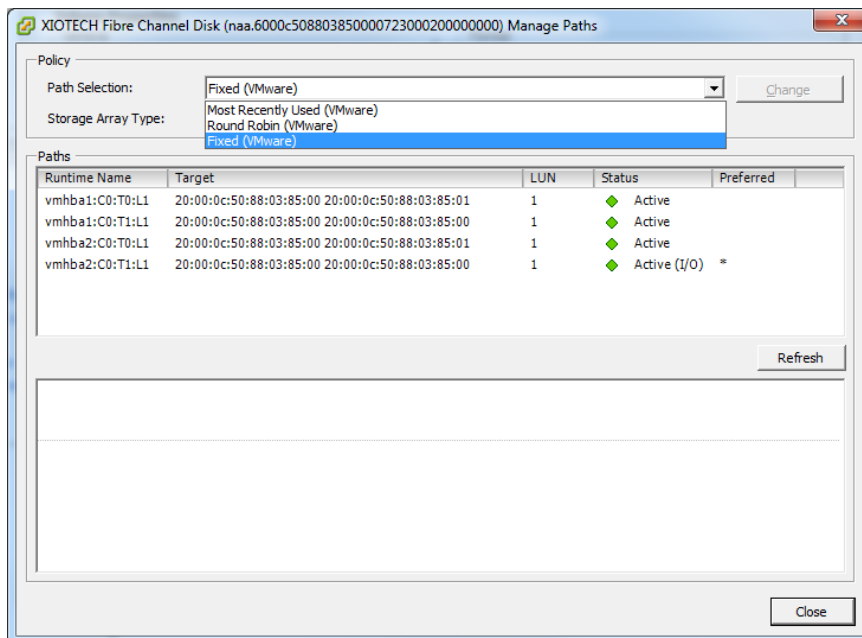## VMware General Recommendations

Some general guidelines for optimal use with VMware ESXi are as follows:

- Preferred allocation size settings (Windows VMs)
- Permanent Device Loss (PDL) Setting
    In vSphere Client, select the host
    1. Click the **Configuration** tab, then click **Advanced Settings** under Software
    2. Click **Disk** in the left panel
    3. Set **Disk.AutoremoveOnPDL** to **0** then click OK

- Round-Robin PSP (Path Selection Policy) for native multipath (NMP) settings for QV2020 volumes is recommended, *except for the situations described in VMware's:* Setup for Failover Clustering and Microsoft Cluster Service.

    Round Robin optimal performance can be gained by decreasing the VMware IOPS default (1000) to an industry recommendation of 1, for SSD and NVMe devices. This can be done by creating a rule (for all devices from a QV2020) or by changing the setting on a device. Creating a rule prior will allow for all presented devices from the QV2020 to be marked for Round Robin. Below is the rule:

esxcli storage nmp satp rule add -s "VMW_SATP_DEFAULT_AA" -V "XIOTECH" -M "ISE4400" -P "VMW_PSP_RR" -O "iops=1"

(where **-M is ISE4400** is model for QV2020 or QV1020 FC **or ISE4401** is model for QV2020 and QV1020 iSCSII)

- Fixed PSP (Path Selection Policy) is also supported, and selectable in VMware as shown:

ESXi 6.5 and ESXi 6.7

Caution: VMware support has ended for **ESXi 5.5 and 6.0 and 6.1**. *This environment is no longer supported by VIOLIN, as well. Use at your own risk. Recommendations previously provided prior to the end of support are carried forward in the below sections, for reference. Future interoperability testing is not guaranteed by VIOLIN.*

Refer to "Fibre Channel HBA Settings" section to ensure that the HBA parameters are set correctly. Ensure that the latest ESXi updates are also applied.

A reboot of the ESXi host will most likely be required when changing these parameters. If you are using a VMware HA configuration, cluster nodes can be rebooted one at a time. The assumption is that virtual machines are migrated to other nodes while doing maintenance on a given cluster node. Verify the changes have been made and then migrate virtual machines back to their preferred nodes if needed after the reboot.

Note: ESXi 6.5 requires the latest updates to obtain path recovery fixes and fixes specific to UCS. The latest fixes for UCS are in Dec 2017.

Direct-Attached Login Timeout Values for QLogic HBAs

To facilitate faster logins following link down or link up events in direct-attached configurations with Array, the following `ql2xlogintimeout` parameter for QLogic HBAs is recommended for use with VMware ESXi 5.5, ESXi 6.0, ESXi 6.5 and ESXi 6.7.

```
# esxcli system module parameters set -p

ql2xlogintimeout=5 -m qla2xxx
```

Following the host reboot, use the following guidelines to check the `ql2xlogintimeout` parameter and ensure that it is set properly:

```
# esxcli system module parameters list -m qla2xxx |

grep ql2xlogintimeout ql2xlogintimeout int 5 Login

timeoutvalue in seconds.
```

HBA Queue Depth Considerations

Review the VMware Knowledge Base references that relate to setting HBA queue depth and configuring VM disk request limits in attempting to increase queue depth or throttling the queue depth down to prevent overloading the storage.

Change Maximum Outstanding Disk Requests for Virtual Machines

If you adjusted the LUN queue depth, change the `Disk.SchedNumReqOutstanding` parameter so that its value matches the queue depth. The parameter controls the maximum number of outstanding requests that all virtual machines can issue to the LUN. Change this parameter only when you have multiple virtual machines active on a LUN. The parameter does not apply when only one virtual machine is active. In that case, the bandwidth is controlled by the queue depth of the storage adapter.

Procedure

1. In the vSphere Client, select the host in the inventory panel.

2. Click the **Configuration** tab and click **Advanced Settings** under Software.
3. Click **Disk** in the left panel and scroll down to **Disk.SchedNumReqOutstanding**.
4. Change the parameter value to the number of your choice and click **OK**.

This change can impact disk bandwidth scheduling, but experiments have shown improvements for disk intensive workloads.

Raw Device Maps (RDMs)

When using RDMs with VMware ESXi 5.5, ESXi 6.0 and ESXi 6.5, the SCSI timeouts on RDMs should be adjusted on Linux Guest OS versions. The default SCSI timeouts for RDMs on Linux Guest OS versions are set to **30 seconds**. VIOLIN recommends setting this value to at **least 120 seconds** to cover certain conditions, such as controller firmware upgrades for the QV2020 platform.

To change these values, the administrator can create a script on the Linux Guest OS under the **/etc/udev/ rules.d** directory. As an example, the script in use is called **99-scsi-disk.rules**. This does not affect virtual disk timeout values, which are set to 180 seconds by default.

```
# cat 99-scsi-disk.rules

#Addition by USER to test RDM scsi timeout settings.

ACTION=="add", SUBSYSTEMS=="scsi",

SYSFS{type}=="0|7|14", RUN+="/bin/sh -c 'echo 120 >

/sys/$DEVPATH/timeout'"
```

A Linux Guest OS with existing RDMs may need to be rebooted for this to take effect.

RDMs look like a normal physical disk when viewed from within the Guest OS with the **/usr/bin/lsscsi** command:

```
# lsscsi

[1:0:0:0] cd/dvd NECVMWar VMware IDE CDR10 1.00

/dev/sr0

[2:0:0:0] disk VMware Virtual disk 1.0

/dev/sda

[2:0:1:0] disk XIOTECH ISE4400 A

/dev/sdb

[2:0:2:0] disk XIOTECH ISE4400 A

/dev/sdc
```

Once the parameter is properly set, it can be verified in the **/sys/class/scsi_disk** directory structure. For example:

```
# cat sys/class/scsi_disk/2:0:1:0/device/timeout

120
```

Performance Tuning

Please refer to the VMware storage queues and performance article reference and the Storage Performance Analysis and Monitoring article. The links may be found in the External References section.

Also, refer to the VMware DiskMaxIOSize document as a way to improve latencies.

Linux

Red Hat Enterprise Linux (RHEL)/CentOS and SUSE Linux Enterprise Server (SLES)

- Red Hat EL 7.2 is the recommended minimum version for 7.x

The below are best practices recommendations for RHEL and SLES.

HBA Configuration

For most fabric-based configurations, the HBA's default settings are sufficient.

For direct-attached systems, we recommend that the HBA's queue depth parameter **ql2xmaxqdepth** be changed to 256 and the login timeout parameter **ql2xlogintimeout** be 5.

The procedures to set these parameters can vary based on HBA and operating system. Please follow the instructions in QLogic's Linux driver documentation from QLogic's web site.

Multipath Settings

All multipath configurations are made in the **multipath.conf** file. This file is located in the **/etc** directory. After changes are made to the **multipath.conf** file, the service needs to be told to re-read it. Please refer to the vendor operating system documents for information on how to do this.

Verify that the polling interval is set to 10 in the defaults section of the **/etc/multipath.conf** file, as follows:

```
defaults {
            polling_interval 10
}
```

Note: There may be additional settings in the defaults section of your multipath.conf file depending on the requirements of your fabric.

Device-specific modifications must be made in the Devices section of the **/etc/multipath.conf** file. Refer to the RHEL or SLES documentation for the chosen multipath product for additional information on these parameters.

The following are examples of what the device entries would look like in the **multipath.conf** file:

Note: Input the correct model number for the product that you are using: **ISE4400** for Fibre or **ISE4401** for iSCSI

For the latest 2 updates of Redhat 7.x/CentOS

```
Defaults     {
            polling_interval    10
}            user_friendly_namesyes
Devices      {
```

```
                    Vendor              "XIOTECH"

                    Product             ISE4400

                    path_group_policy   multibus

                    prio_callout        "none"

                    path_checker        tur

                    path_selector       "round-robin  0"

                    Failback            immediate

                    dev_loss_tmo        600

                    fast_io_fail_tmo    5

                    no_path_retry       12

                    rr_min_io_req       1

        }
```

For SLES11 SP3/SP4

```
    Defaults        {

                    polling_interval    10

    }               user_friendly_names no

    Devices         {

                    device                  {

                    vendor                  "XIOTECH"

                    product                 ISE4400

                    path_grouping_policymultibus

                    path_selector           round-robin 0"

                    uuid_attribute          "ID_SERIAL"

                    path_checker            tur

                    rr_min_io_req           1

                    rr_weight               priorities

                    failback                immediate

                    no_path_retry           12

                    dev_loss_tmo            600

                    fast_io_fail_tmo        5

        }
```

Note: Older version of Redhat and SLES may not have the option to set `fast_io_fail_tmo` or `dev_loss_tmo` via the multipathd daemond. In these cases, a script may need to be written by the server's administrator that will manually set the parameters on boot.

After `multipathd` has been reloaded with the configuration from your **multipath.conf** file, present a new LUN to the host and check that the host sees the LUN.

Example (in this case we have 8 paths):

```
[root]# multipath -ll

mpatha (36001f931033a00000226000200000000) dm-1 XIOTECH,ISE4400 size=40G
features='1 queue_if_no_path' hwhandler='0' wp=rw
 `-+- policy='round-robin 0' prio=1 status=active
   |- 2:0:6:0    sdd 8:48     active ready    running
   |- 3:0:4:0    sdo 8:224    active ready    running
   |- 2:0:4:0    sda 8:0      active ready    running
   |- 3:0:6:0    sdr 65:16    active ready    running
   |- 2:0:9:0    sdi 8:128    active ready    running
   |- 3:0:8:0    sdu 65:64    active ready    running
   |- 2:0:12:0   sdl 8:176    active ready    running
 `- 3:0:9:0      sdv 65:80    active ready    running
[root]#
```

And, check for the new values in **/sys/class/fc_remote_ports/rport-<pci>.**

Example:

```
[root]# pwd

/sys/class/fc_remote_ports

[root]#ls

rport-2:0-0    rport-2:0-14   rport-2:0-7    rport-3:0-11   rport-3:0-4

rport-2:0-1    rport-2:0-2    rport-2:0-8    rport-3:0-12   rport-3:0-5

rport-2:0-10   rport-2:0-3    rport-2:0-9    rport-3:0-13   rport-3:0-6

rport-2:0-11   rport-2:0-4    rport-3:0-0    rport-3:0-14   rport-3:0-7

rport-2:0-12   rport-2:0-5    rport-3:0-1    rport-3:0-2    rport-3:0-8

rport-2:0-13   rport-2:0-6    rport-3:0-10   rport-3:0-3    rport-3:0-9


[root]#cd rport-3\:0-8/
[root]# ls
device maxframe_size port_name roles supported_classes dev_loss_tmo node_name
port_state scsi_target_id uevent fast_io_fail_tmo port_id power subsystem
[root]#
[root]# cat dev_loss_tmo
6000
[root]# cat fast_io_fail_tmo
5
[root]#
```

AIX

AIX is supported on the QV2020 and QV1020 Fibre Channel Arrays for:

- AIX 7.1 7100-02 minimum
- AIX 7.2 7200-05 minimum

QV2020 and QV1020 Multipath driver is required, with an IBM OEM HBA.

Multipath Driver

The installation kit provided by Violin System's for support with AIX multipathing is required.

The kit contains:

- ViolinSystems-install-MPIO - script which installs the MPIO files
- ViolinSystems-uninstall-MPIO - removes VS20 entries from the ODM
- vs20.msg - message files from which catalog is derived
- VS20.template - ODM device entries for the ViolinSystems

Installing Multipath

Prior to presenting volumes, it is required that the following attributes be changed on the IBM FC adapter. These can be changed via the **smit fastpath "fscsidd" or using chdev:**

- ➤ Dynamic Tracking of FC Devices must be set to "yes"

    get list of FC devices>

    lscfg -v -l fscsi*

    set per device>

    chdev -l <DEVICE> -a dyntrk=yes -P

- ➤ FC Fabric Event Error Recovery Policy must be set to "fast_fail"

    get list of FC devices>

    lscfg -v -l fscsi*

    set per device>

    chdev -l <DEVICE> -a fc_err_recov=fast_fail -P

- ➤ If changing all FC adapters in the system you can run

lscfg -v -l fscsi* | awk {'print$1'} | xargs -I {} chdev -l {} -a dyntrk=yes -a fc_err_recov=fast_fail -P

If any child devices are present on the adapters, they must be removed prior to setting these attributes.

Best Practices and Configuration Guide

1.   Run the "ViolinSystems-install-MPIO" script. "ViolinSystems-install-MPIO -help" will give usage instructions.
2.   Removal of ViolinSystems AIX Multipath
3.   Run the ViolinSystems-uninstall-MPIO script to remove the ODM entries created by install.

Troubleshooting

Installation Errors:

During install the following message may appear:

> 0518-305 odmadd Could not open object class sm_cmd_hdr
>
> > Check path name and permissions.

This message occurs when the links from /etc/objrepos to /usr/lib/objrepos,

for the following objects, are missing.

> * sm_cmd_hdr
>
> * sm_cmd_hdr.vc
>
> * sm_cmd_opt
>
> * sm_cmd_opt.vc
>
> * sm_menu_opt
>
> * sm_menu_opt.vc
>
> * sm_name_hdr
>
> * sm_name_hdr.vc

If this occurs recreate the links using the "ln" command.

Luns Already Discovered

 If ViolinSystem LUNs have already been discovered and are listed as type "fc other", they can be removed using:

> "rmdev -dl hdisk<N>"
>
> And rediscovered by running "cfgmgr" or by simply rebooting the server.

Windows Server 2008 & 2008 R2: *no longer supported*

Caution: Microsoft ended support of Windows 2008 on 14-Jan-2020. This environment is no longer supported by VIOLIN, as well. Use at your own risk. Below are recommendations previously provided prior to the end of support. Future interoperability testing is not guaranteed by VIOLIN.

VIOLIN recommends the use of Windows Native MPIO for QV2020 storage.

Native multipathing on Windows 2008 and Windows 2008 R2 can generate delays in path discovery of LUNs during QV2020 Upgrade. These delays can result in applications aborting or boot from SAN environments hanging. **A maintenance window is recommended** *for QV2020 firmware upgrades.*

Windows Server 2012 R2 and Windows Server 2016

Multipathing

VIOLIN recommends the use of Windows Native MPIO to the Array LUNs. In order to improve the availability of the storage during common path loss scenarios, including firmware upgrade, the following MPIO DSM registry key changes are recommended.

Note: Values in this table are **decimal**, not hex.

Parameter Name	Default Value	VIOLIN Systems Value
PathVerifyEnabled	0	0
PathVerificationPeriod	30	30
PDORemovePeriod	20	50
RetryCount	3	10
RetryInterval	1	5
UseCustomPathRecoveryInterval	0	1
PathRecovryInterval	40	25
DiskPathCheckDisabled	1	0
DiskPathCheckInterval	10	25

HKLM /System/CurrentControlSet/Services/MPIO/Parameters

Claiming MPIO disks causes the host to reboot. To configure native MPIO to claim Array volumes, use the command lines below. Note that this assumes that the MPIO feature is already installed and that the items in the quotes are case-sensitive.

```
mpclaim -r -i -d "XIOTECH ISE4400"
```

Required Hot-Fixes for Windows 2012 R2

- Microsoft Knowledge Base Article 2779768
- Microsoft Knowledge Base Article 2784160
- Microsoft Knowledge Base Article 2870270
- Microsoft Knowledge Base Article 2803748
- Microsoft Knowledge Base Article 2770917

- Microsoft Knowledge Base Article 976424
- Microsoft Knowledge Base Article 2784679

Windows format on Dedupe Volumes

When formatting a Dedupe volume on Windows using the NTFS filesystem, the Allocation unit size should be set to 8192 for optimal performance.
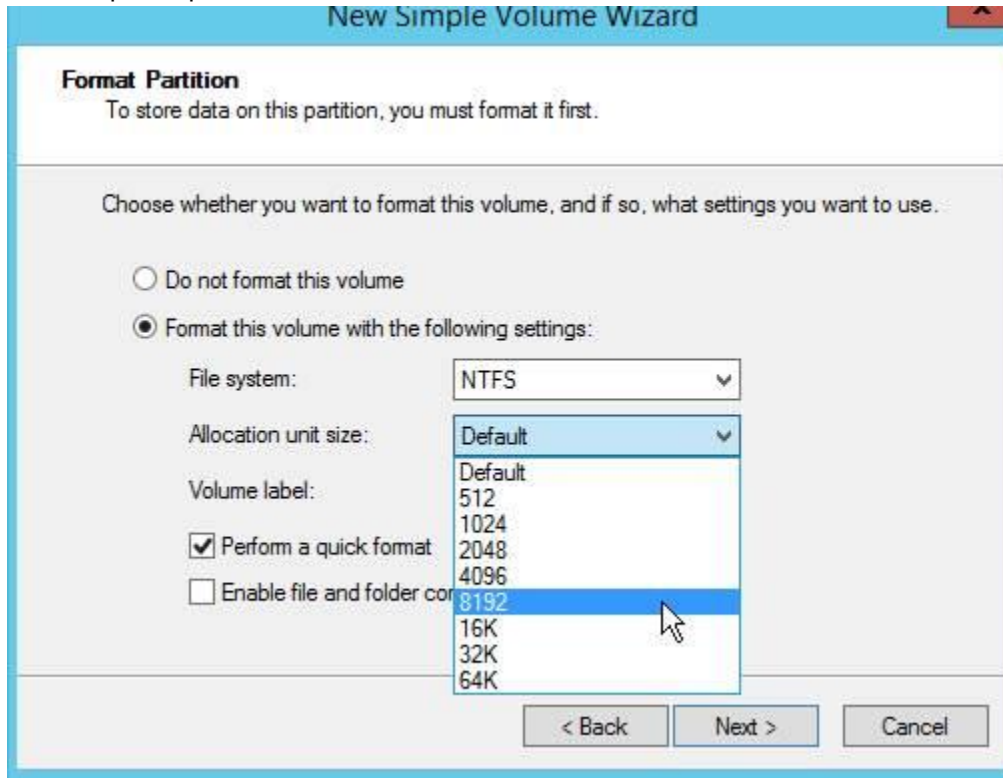

Hyper-V and CSV Settings

Due to the way cluster shared volumes (CSVs) work, it is strongly recommended to not place CSV volumes into maintenance mode as this will have the effect of taking all roles on that CSV offline. Instead of using maintenance mode for CSVs, it is recommended to disable heartbeat checking for the duration of Array software upgrades.

To disable virtual machine heartbeat checking, perform the following steps on each Hyper-V virtual machine role:

1. Open the Failover Cluster Manager.
2. Connect to the cluster (if not connected automatically).
3. Locate each virtual machine role in the cluster.
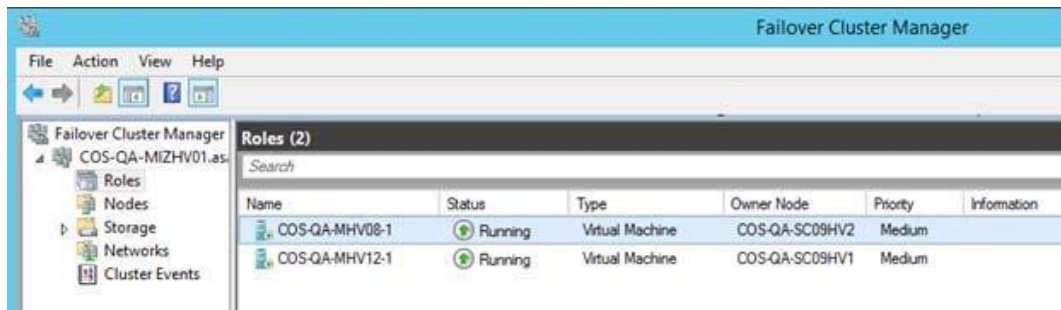
4. Select each VM, one at a time.

5. In the details pane (lower center section of the failover cluster manager) click on the **Resources** tab.
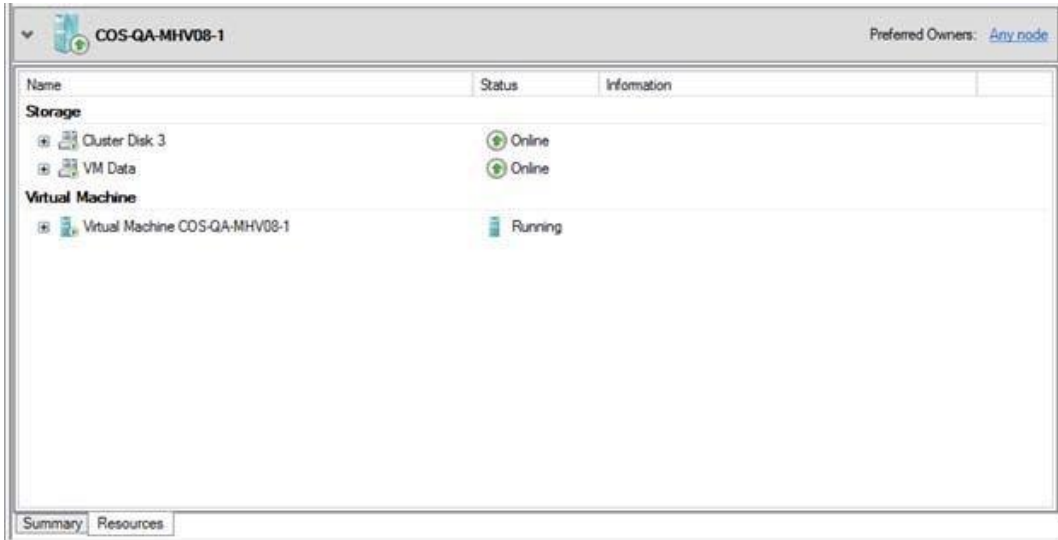
6. Right-click the Virtual Machine {VM name} section and then select **Properties**.
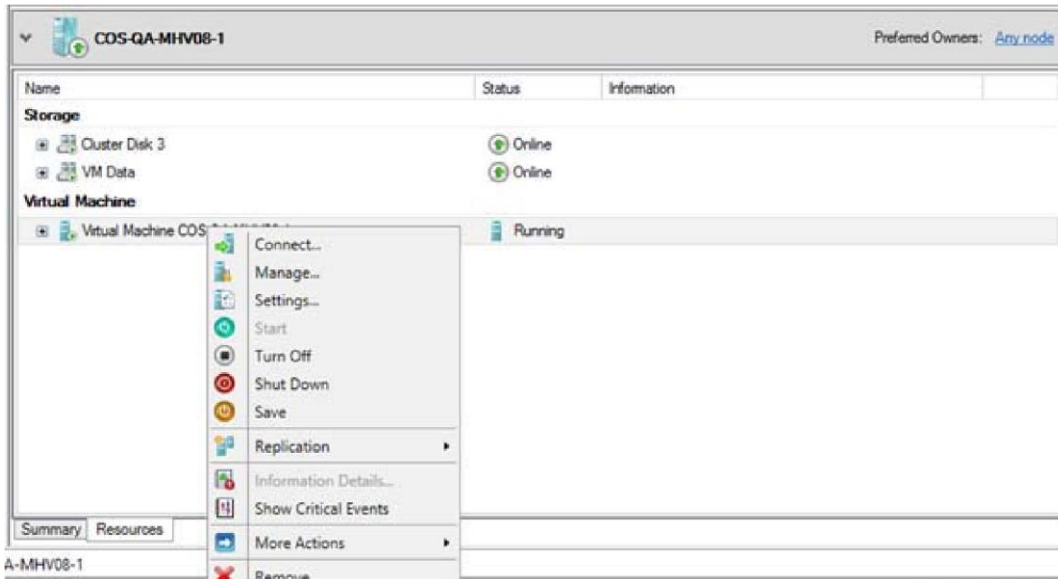
7. Select the **Settings** tab, and then clear the check boxes for *Enable heartbeat monitoring for the virtual machine*.
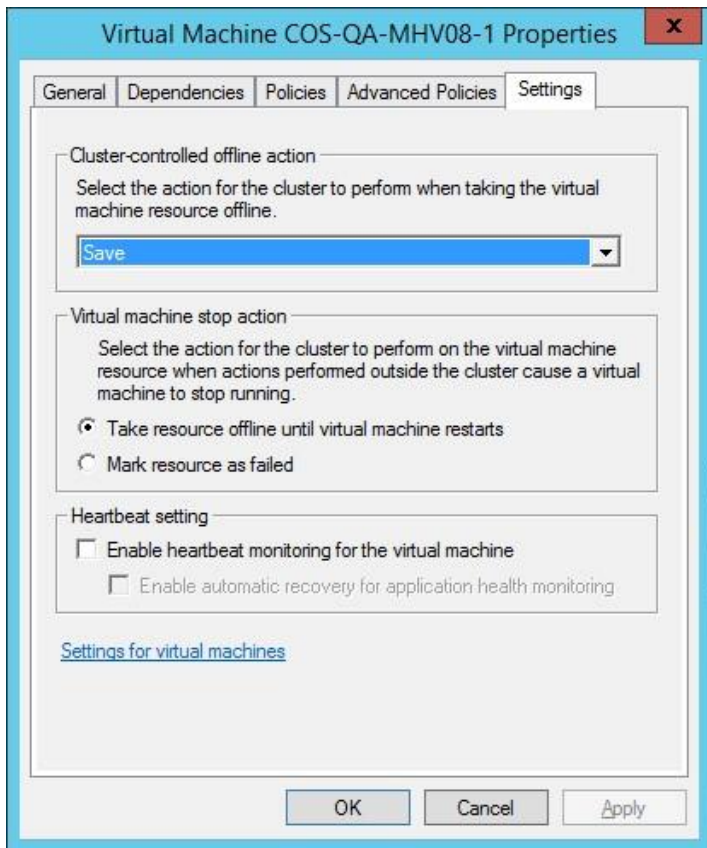
8. Ensure that the Cluster-controlled offline action is set to Save.
9. Ensure that the Virtual Machine stop action is set to Take resource offline until virtual machine restarts.
10. Click **OK**.

After the Array software upgrade process is completed, these changes can be reverted to the customer settings.

Array Software Upgrade Options

There are some differences between cluster nodes and standalone servers when it comes to upgrading the Array. Check "Online Upgrade (clusters)" below to perform an online upgrade.

Online Upgrade (standalone servers)

-   All hot fixes listed must be installed.

Offline Upgrade

-   All hot fixes listed must be installed.

Online Upgrade (clusters)

- All hot fixes listed must be installed.
- No boot from SAN volumes—Some issues have been reported that seem to occur only on cluster nodes booting from SAN. Until the cause is determined and a workaround or a fix is available, VIOLIN recommends performing offline upgrades with cluster nodes configured to boot from SAN.
  - o "Cluster Service Resource Maximum Restart" count set to be equal to three times the number of Fibre Channel Arrays + 1. As an example, if there are four Arrays that contribute storage to the cluster, the restart count would be set to 13 ((3 * 4) + 1). This is required for the duration of the upgrade process—can be set to other values for normal operations.
- VIOLIN recommends using the cluster maintenance mode to perform upgrades.

Configuring Cluster Disk Resource Failover Policies

Follow the steps below to configure the cluster resource failover policies.

1. Open the Failover Cluster Manager GUI.
2. Select the Storage node in the left-hand navigation pane.
3. Select an Array clustered disk, right-click, and then select **Properties** from the drop-down menu.
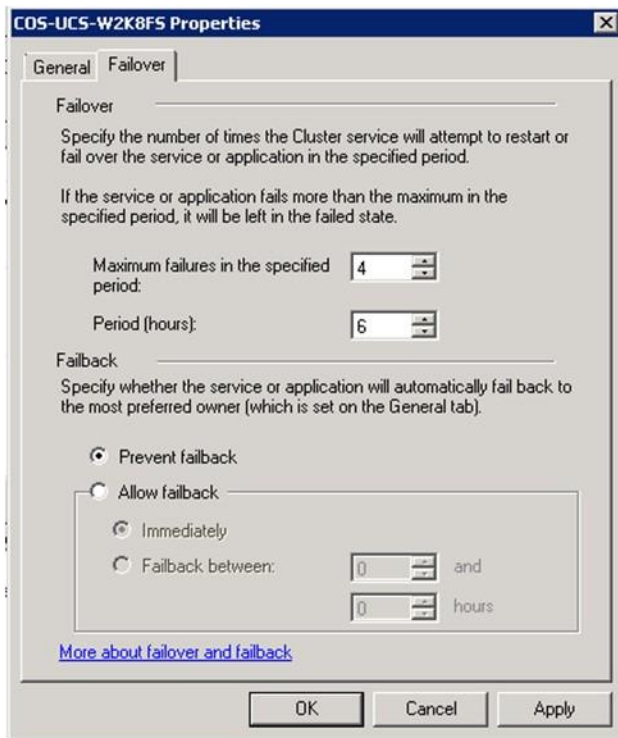
Cluster Disk Properties

---

4. Select the **Policies** tab.

5. Change the **Maximum restarts in the specified period** value. This should be equal to 1 + three times the number of Arrays in the cluster. In the above, the value is set to 4, which is appropriate for a single Array.
6. Repeat for each Array disk in the cluster.

Configuring Clustered Services and Applications

1. Expand the **Services and Applications** node in the left-hand navigation pane.
2. Select each service or application, right-click, and then select **Properties** from the drop-down.
3. Select the **Failover** tab.

4. Change the **Maximum failures in the specified period.** This value should be equal to 1 + three times the number of Fibre Channel Arrays in the cluster.

5. In the main pane for each clustered application, set the failover properties for each resource.

Clustered Application Resource List

6. Right-click each resource in the resource list, select **Properties**, and then select the **Policies** tab.
7. Change the Maximum restarts in the specified period to the same value used for the cluster disk resource.
8. Repeat these steps for each resource in each clustered service or application.

Setting Maintenance Mode During an Upgrade

With the exception of Clustered Shared Volumes (CSVs), VIOLIN strongly recommends placing clustered storage into maintenance mode, which prevents failover of cluster disks while running Array software upgrades. See "Linux" on page 43 for details on configuring Virtual Machines on CSVs. The following process demonstrates how to enable and then disable maintenance mode.

1. Open Failover Cluster Manager.
2. Navigate to the **Storage** node in the left-hand navigation pane.
3. Right-click each Array volume, select **More Actions**, then select **Turn On Maintenance Mode for this disk**.
4. The GUI shows that the disk is in maintenance mode.

COS-UCS-W2K8FS

⊟ ▭ Cluster Disk 2	🔼 Online (Maintenance)	COS-UCS-W2K8-n2
Volume: (E)	File System: NTFS	10 GB (99.2% free)

Disk Maintenance Mode

5. Once all Fibre Channel Arrays are upgraded, turn off maintenance mode using the same process (select **Turn Off Maintenance Mode for this disk**).

In order to improve the availability of the storage during common path loss scenarios, including firmware upgrade, the following MPIO DSM registry key changes are recommended.

Note: Values in this table are decimal, not hex.

Parameter Name	Default Value	VIOLIN Systems Value
PathVerifyEnabled	0	0
PathVerificationPeriod	30	30
PDORemovePeriod	20	50
RetryCount	3	10
RetryInterval	1	5
UseCustomPathRecoveryInterval	0	1
PathRecovryInterval	40	25
DiskPathCheckDisabled	1	0

DiskPathCheckInterval	10	25

HKLM /System/CurrentControlSet/Services/MPIO/Parameters

Claiming MPIO disks causes the host to reboot. To configure native MPIO to claim the Legacy QV2020 and Fibre Channel Array volumes, use the command lines below. Note that this assumes that the MPIO feature is already installed and that the items in the quotes are case-sensitive.

```
mpclaim -r -i -d "XIOTECH Array4400"
```

## Disabling FUA for QV2020 and QV1020 Volumes

Windows Server operating systems have a mechanism that writes directly to the disk and instructs the disk to not use any write caching mechanisms it may have. This uses part of the SCSI block commands known as "Force Unit Access," or FUA. This is done to protect against data loss on storage devices (USB hard drives, etc.) that do not have a battery-backed caching capability. The QV2020 and QV1020 support write caching and have a battery backup system to preserve that cache, so performance can be increased by disabling this mechanism.
Open Disk Manager.

1. Right-click the Array LUN (Fibre Channel Array).
2. Select **Properties**.
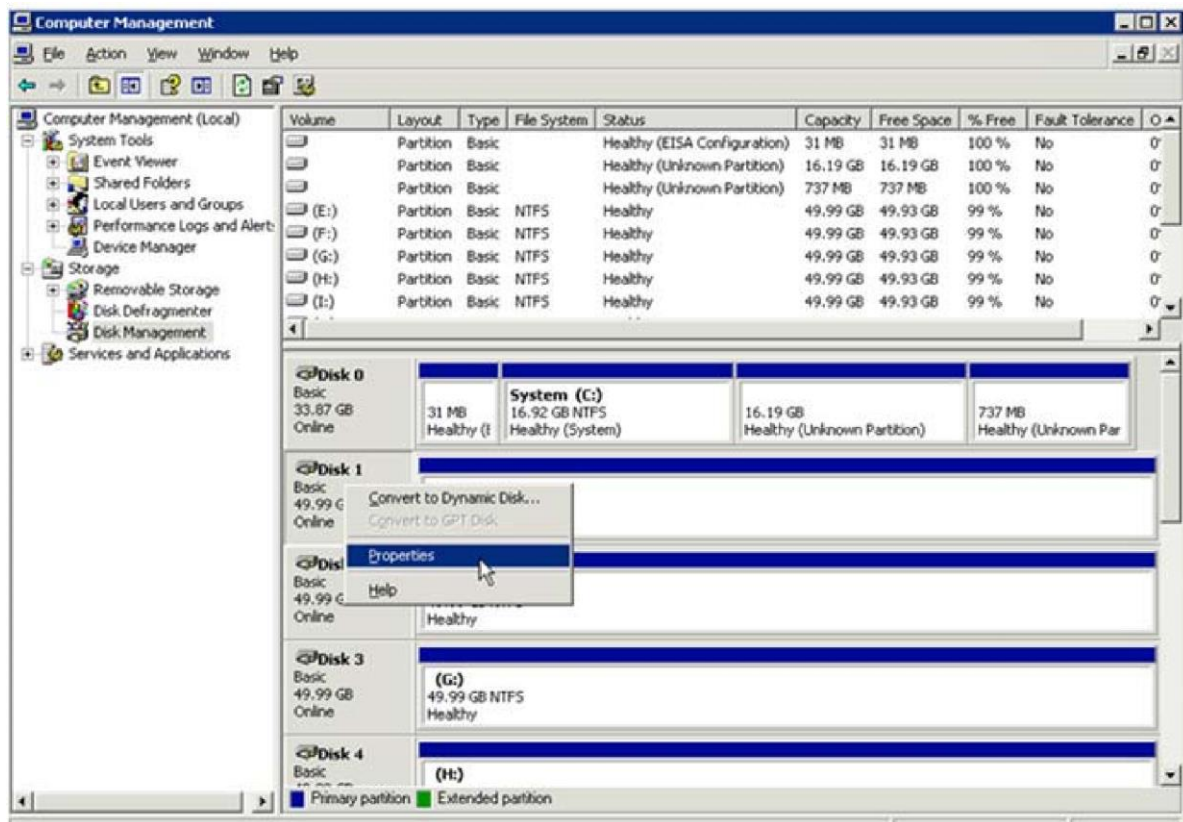
3. On the *Disk Device Properties* dialog, select the **Policies** tab.
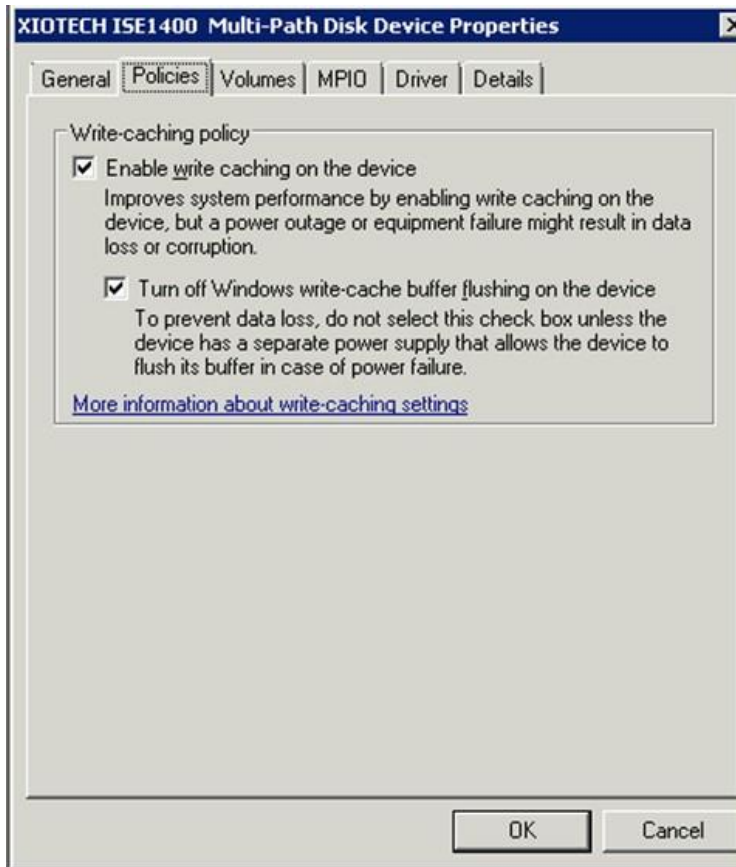
4. Ensure that the Enable write caching on the disk and Enable advanced performance check boxes are selected.
5. Click **OK**.
6. Repeat for each Array volume.

Required Hot Fixes

Required Hot-Fixes for Windows 2008 SP2

- Microsoft Knowledge Base Article ID 968675
- Microsoft Knowledge Base Article ID 970525
- Microsoft Knowledge Base Article ID 972797
- Microsoft Knowledge Base Article ID 974878
- Microsoft Knowledge Base Article ID 976674
- Microsoft Knowledge Base Article ID 976748
- Microsoft Knowledge Base Article ID 977001
- Microsoft Knowledge Base Article ID 977153
- Microsoft Knowledge Base Article ID 977287
- Microsoft Knowledge Base Article ID 977675
- Microsoft Knowledge Base Article ID 977890
- Microsoft Knowledge Base Article ID 978157
- Microsoft Knowledge Base Article ID 979458
- Microsoft Knowledge Base Article ID 979743
- Microsoft Knowledge Base Article ID 979764
- Microsoft Knowledge Base Article ID 981357

- Microsoft Knowledge Base Article ID 2406705
- Microsoft Knowledge Base Article ID 2522766

Windows format on Dedupe Volumes

When formatting a Dedupe volume on Windows using the NTFS filesystem, the Allocation unit size should be set to 8192 for optimal performance.
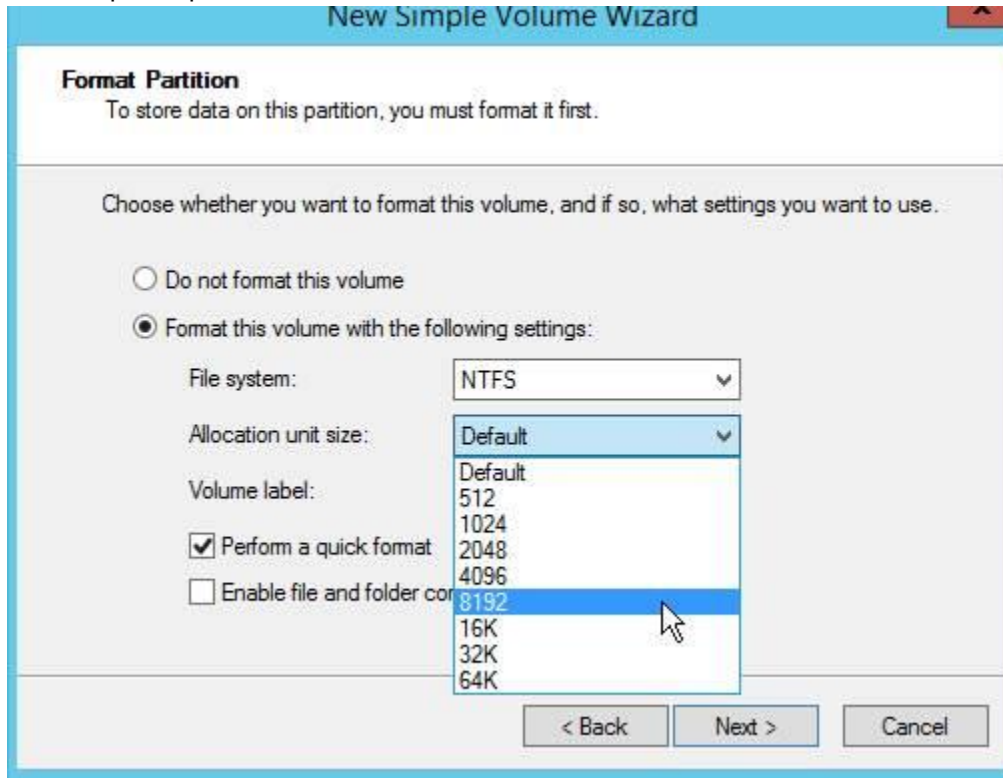
QV2020 and QV1020 Software Upgrade Options

There are some differences between cluster nodes and standalone servers when it comes to upgrading the Array. Check "Online Upgrade (clusters)" below to perform an online upgrade.

*Caution: Native multipathing on Windows 2008 and Windows 2008 R2 can generate delays in path discovery of LUNs during Array Upgrade. These delays can result in applications aborting or boot from SAN environments hanging. A **maintenance window is recommended** for this version of Windows.*

Online Upgrade (standalone servers)

- All hot fixes listed must be installed.

Offline Upgrade
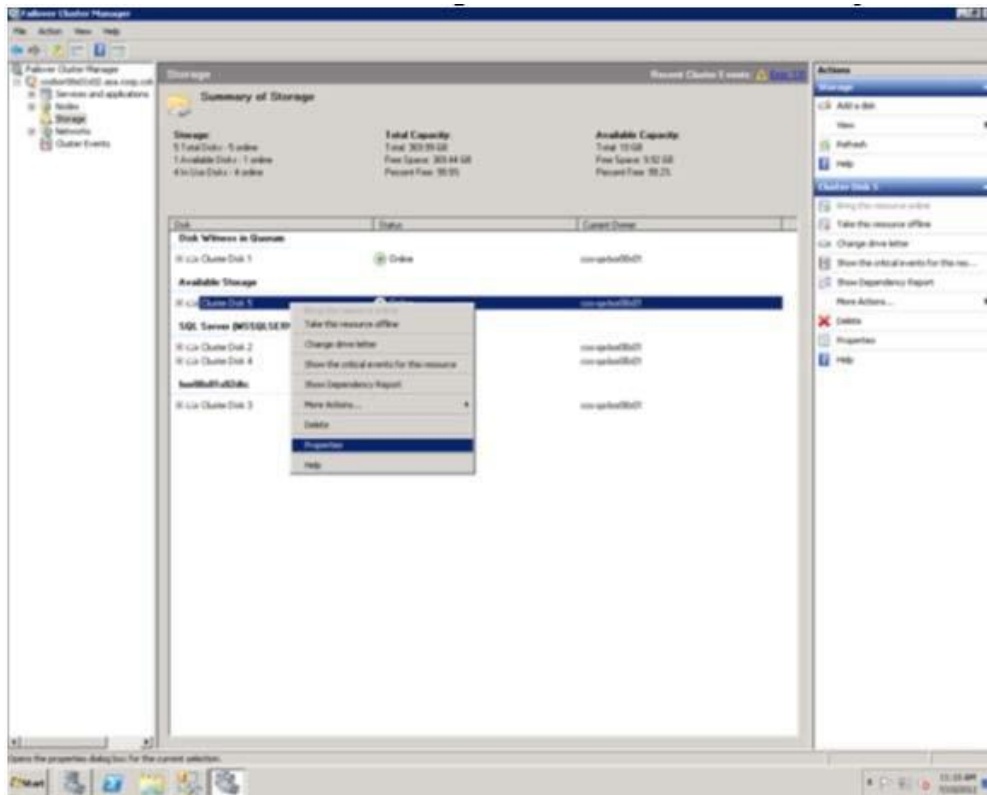
- All hot fixes listed must be installed.

Online Upgrade (clusters)

- All hot fixes listed must be installed.
- **No boot from SAN volumes**—Some issues that seem to occur only on cluster nodes booting from SAN have been reported. Until the cause is determined and a workaround or a fix is available, VIOLIN recommends performing offline upgrades with cluster nodes configured to boot from SAN.
- "Cluster Service Resource Maximum Restart" count set to be equal to three times the number of Fibre Channel Arrays + 1. As an example, if there are four Arrays that contribute storage to the cluster, the restart count would be set to 13 ((3 * 4) + 1). This is required for the duration of the upgrade process—can be set to other values for normal operations.
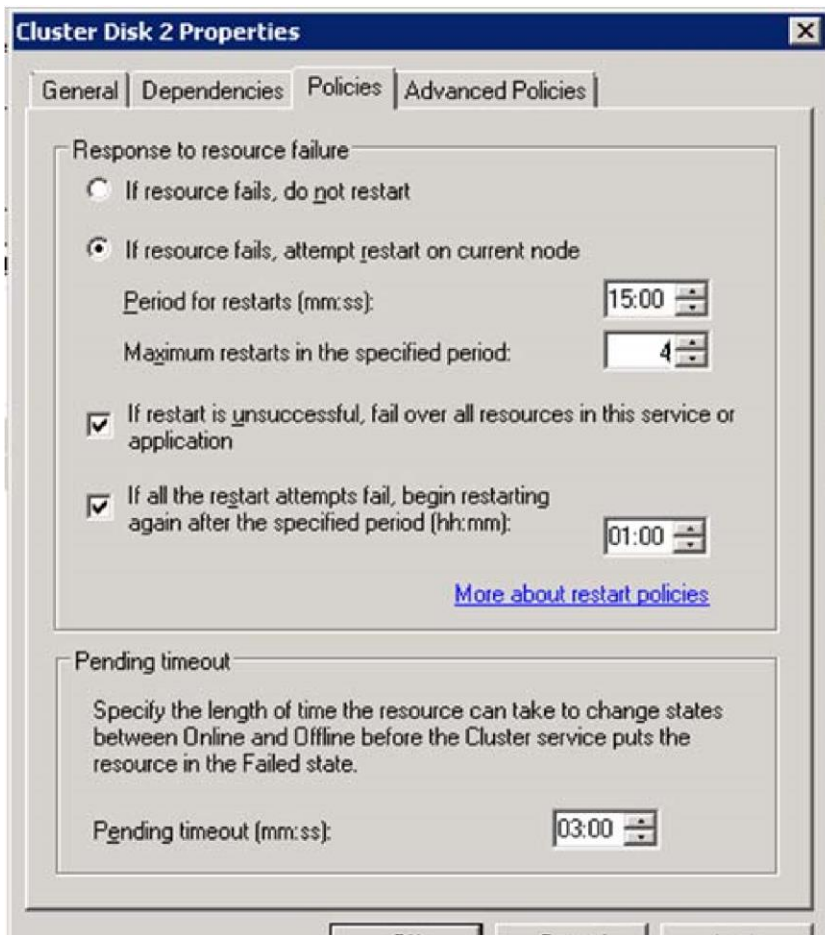- VIOLIN recommends using the cluster maintenance mode to perform upgrades.

Configuring Cluster Disk Resource Failover Policies

Follow the steps below to configure the cluster resource failover policies.

1. Open the Failover Cluster Manager GUI.
2. Select the Storage node in the left-hand navigation pane.
3. Select an Array clustered disk, right-click, and then select **Properties** from the drop-down menu.
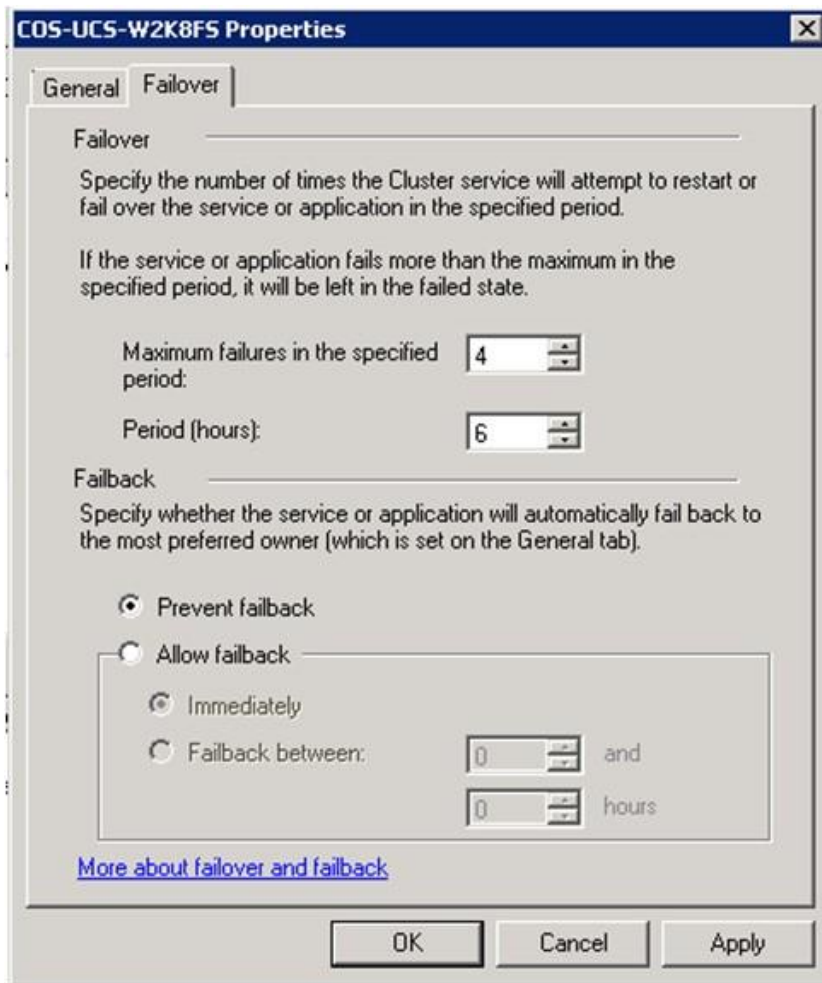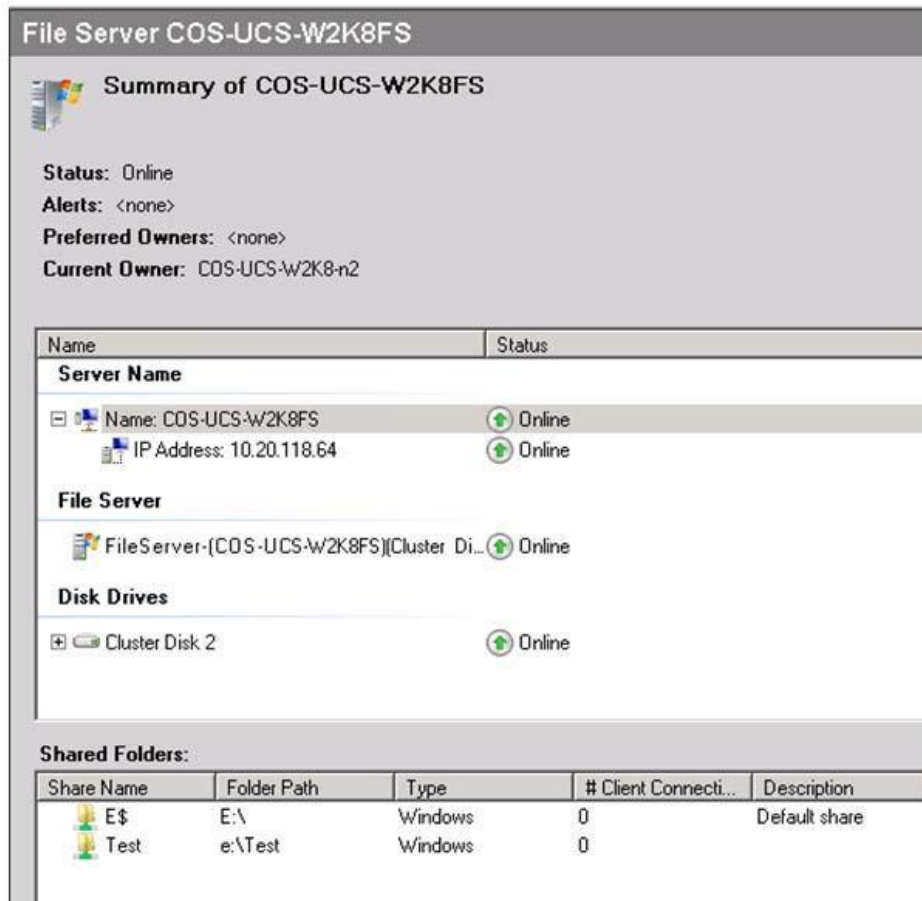

4. Select the **Policies** tab.

5. Change the Maximum restarts in the specified period value. This should be equal to 1 + three times the number of Arrays in the cluster. In the above, the value is set to 4, which is appropriate for a single Array.
6. Repeat for each Array disk in the cluster.

Configuring Clustered Services and Applications

1. Expand the **Services and Applications** node in the left-hand navigation pane.
2. Select each service or application, right-click, and then select **Properties** from the drop-down.
3. Select the **Failover** tab.

4. Change the Maximum failures in the specified period. This value should be equal to 1 + three times the number of Fibre Channel Arrays in the cluster.
5. In the main pane for each clustered application, set the failover properties for each resource.

6. Right-click each resource in the resource list, select **Properties**, and then select the **Policies** tab.
7. Change the Maximum restarts in the specified period to the same value used for the cluster disk resource.
8. Repeat these steps for each resource in each clustered service or application.

Setting Maintenance Mode During an Upgrade

VIOLIN strongly recommends placing clustered storage into maintenance mode, which prevents failover of cluster disks while running Array software upgrades. The following process demonstrates how to enable and then disable maintenance mode.

1. Open Failover Cluster Manager.
2. Navigate to the **Storage** node in the left-hand navigation pane.
3. Right-click each Array volume, select **More Actions**, then select **Turn On Maintenance Mode for this disk**.
4. The GUI shows that the disk is in maintenance mode.
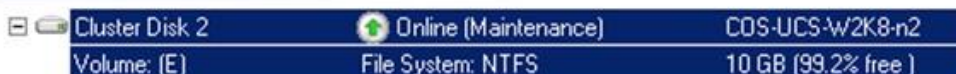
5. Once all Fibre Channel Arrays are upgraded, turn off maintenance mode using the same process (select **Turn Off Maintenance Mode for this disk**).

Using QV2020 and QV1020 with DataCore

For HA and SMPA configurations, VIOLIN Systems recommends using the DataCore HBA drivers and DataCore MPIO.

Upgrade Notes

HA Environments

The DataCore server that is servicing I/O to the Array to be upgraded should be shut down prior to the upgrade. Online upgrade is supported because the other server will continue to service I/O.

SMPA Configurations

Online upgrade of Fibre Channel Array software is not available with SMPA configurations.

Online upgrade of Array software is supported with SMPA and HA configurations.

External References

Storage and Networking Industry Association (SNIA) Dictionary

http://www.snia.org/education/dictionary

IBM Fibre Cable Ratings

https://www.ibm.com/developerworks/mydeveloperworks/blogs/anthonyv/entry/
don_t_say_green_say_aqua1?lang=en

VMware Queue Depth Considerations

http://kb.VMware.com/selfservice/microsites/search.do?language=en_US&cmd=displayKC&externalId=1267

VMware Maximum Disk Requests per VM

http://kb.VMware.com/selfservice/microsites/search.do?language=en_US&cmd=displayKC&externalId=1268

Red Hat Enterprise QV2020 Linux (RHEL) MPIO Documentation

http://docs.redhat.com/docs/en-US/Red_Hat_EnterprArray_Linux/index.html

SUSE Linux Enterprise QV2020 Server (SLES) MPIO Documentation

http://www.novell.com/documentation

QLogic Red Hat Enterprise QV2020 Linux 6.x Driver

http://filedownloads.qlogic.com/files/Driver/81208/readme_FC-FCoE_Inbox_driver_update.txt

VMware Storage Queues and Performance

http://communities.vmware.com/docs/DOC-6490

VMware Storage Performance Analysis and Monitoring

http://communities.vmware.com/docs/DOC-5490

VMware Disk MaxIOSize

http://kb.vmware.com/kb/1003469

VMware iSCSI Design and Considerations Guide—

http://www.vmware.com/resources/techresources/1006

VIOLIN

Better Business Performance

2375 Telstar Drive #150
Colorado Springs, CO 80920
United States of America

www.violinsystems.com