Whitepaper

Congestion Spreading and How to Avoid It

This document describes how Congestion Spreading (a.k.a., Slow Drain) can impact your Storage Area Network (SAN), the metrics used to describe the severity of each type of congestion for both Connectrix B-Series and the MDS Series as well as preventive measures that can be taken to avoid the effects of Congestion Spreading

May 2019

Congestion Spreading and How to Avoid It | H17762.1 |

\

Revisions

Date	Description
May 2019	Initial release

Acknowledgements

This paper was produced by the following members of the Dell EMC storage engineering team:
Authors:
Alan Rajapa
Erik Smith

The information in this publication is provided "as is." Dell Inc. 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 software described in this publication requires an applicable software license.

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

Table of Contents

1 Preface

This document describes how Congestion Spreading (a.k.a., Slow Drain) can impact your Storage Area Network (SAN), the metrics used to describe the severity of each type of congestion for both Connectrix B-Series and the MDS-Series as well as preventive measures that can be taken to avoid the effects of Congestion Spreading.

As part of an effort to improve and enhance the performance and capabilities of its product line, Dell EMC from time to time releases revisions of its hardware and software. Therefore, some functions described in this document may not be supported by all revisions of the software or hardware currently in use. For the most up-to-date information on product features, refer to your product release notes.

If a product does not function properly or does not function as described in this document, please contact your Dell EMC representative.

Audience

This TechBook is intended for Dell EMC field personnel, including technology consultants, and for the storage architects, administrators, and operators involved in acquiring, managing, operating, or designing a networked storage environment that contains EMC and host devices.

Related Documentation

All related documentation and release notes can be found on https://dell.com/support. Click **Support by Product**, input the product name and click **Documentation**.

EMC Support Matrix and E-Lab Interoperability Navigaton

For the most up-to-date information, always consult the *EMC Support Matrix* (ESM), available through E-Lab Interoperability Navigator (ELN) at http://elabnavigator.EMC.com

Where to get help

Dell EMC support, product, and licensing information can be obtained on the Dell EMC Online Support site as described next.

Note: To open a service request through the Dell EMC Online Support site, you must have a valid support agreement. Contact your Dell EMC sales representative for details about obtaining a valid support agreement or to answer any questions about your account.

Product information

For documentation, release notes, software updates, or information about Dell EMC products, licensing, and service, go to the Dell EMC Online Support site (registration required) at: https://www.dell.com/support

Technical support

Dell EMC offers a variety of support options.

Support by Product — Dell EMC offers consolidated, product-specific information on the Web at: https://support.dell.com/products
The Support by Product web pages offer quick links to Documentation, White Papers, Advisories (such as frequently used Knowledgebase articles), and Downloads, as well as more dynamic content, such as presentations, discussion, relevant Customer Support Forum entries, and a link to Dell EMC Live Chat.

Dell EMC Live Chat — Open a Chat or instant message session with a Dell EMC Support Engineer.

eLicensing support

To activate your entitlements and obtain your license files, visit the Service Center on https://dell.com/support, as directed on your License Authorization Code (LAC) letter e-mailed to you.

2 Overview

The goal of this white paper is to:

1. Describe how Congestion Spreading (a.k.a., Slow Drain) can impact your Storage Area Network (SAN),

2. Define the metrics used to describe each severity and type of congestion for both Connectrix B-Series and MDS-Series products,

3. Describe the preventive measures that can be used to avoid the effects of Congestion Spreading, and

4. Demonstrate how to use the above information to detect, prevent and remediate Congestion Spreading due to oversubscription.

PREREQUISITES

Please Note:

This document assumes the following software versions are in use. The steps may differ in older versions. Please refer to the Appendix for details that describes how to enable the features required.

1. Unisphere for VMAX/PowerMax is installed and running and the array has been registered to collect performance data.
   https://support.emc.com/products/27045_Unisphere-for-/Documentation/?source=promotion

2. SAN Management GUIs are installed.
   a. For Brocade Fabrics: Connectrix Manager Data Center Edition (CMCNE) 14.x or higher
   Download:
   https://support.emc.com/search/?text=CMCNE%2014&searchLang=en_US&facetResource=DOWN
   Admin Guide:
   https://support.emc.com/search/?text=CMCNE%2014%20admin%20guide&searchLang=en_US
   b. For Cisco Fabrics: Cisco Data Center Network Manager(DCNM) 10.x or higher
   Download:
   https://support.emc.com/search/?text=DCNM%2010&facetResource=DOWN

   Admin Guide:
   https://www.cisco.com/c/en/us/support/cloud-systems-management/prime-data-center-network-manager/products-installation-guides-list.html

3. SAN Switch Firmware should be the following:
   a. Brocade: Fabric O.S 7.4.1d or higher
   Download:
   https://support.emc.com/search/?text=Brocade%20FOS%20download&searchLang=en_US&facetResource=DOWN

   b. Cisco: NX-OS 6.2(13) or higher
   Download:
   https://support.emc.com/search/?text=NX-OS%20download

4. All necessary performance monitoring licenses are installed.

   a. Brocade requires a MAPS license: https://docs.broadcom.com/docs/53-1005239-04

   b. Cisco requires DCNM-SAN Server package license: https://www.cisco.com/c/en/us/support/cloud-systems-management/prime-data-center-network-manager/products-installation-guides-list.html

   c. VMAX/PowerMAX requires a Unisphere eLicense. Refer to page 21 of the following PDF for more details: https://www.emc.com/collateral/TechnicalDocument/docu88904.pdf

3 What is Congestion Spreading?

Transporting data to and from a storage array requires all data to be delivered to the destination in a timely fashion. This is especially true for block-based storage protocols that make use of SCSI (e.g. Fibre Channel - FCP). Although the exact reasons for this are outside the scope of this white paper, more detail can be found in the "Congestion and Backpressure" section of the *Networked Storage Concepts and Protocols* Techbook: (https://www.emc.com/collateral/hardware/technical-documentation/h4331-networked-storage-cncpts-prtcls-sol-gde.pdf).

Like any other network protocol, Fibre Channel (FC) needs to ensure this timely delivery of data under a wide range of common network congestion situations. The mechanism used by FC focuses on the prevention of frame loss by using buffer-to-buffer flow control. Because of this, FC is considered a "Lossless Protocol".

Although the flow control mechanisms used by each protocol are slightly different, FC and other Lossless protocols (e.g., DCB Ethernet and inifiniband) prevent buffer overflow at either end of a link by allowing the transmitter to determine when the receiver at other end of the link is nearing capacity. When this determination is made, a port will stop transmitting data until the other end of the link indicates it's ready to receive additional data. While a transmitter is in this state, it's unable to transmit frames and we say that it's experiencing congestion. If a transmitter experiences congestion for a long enough period of time, this congestion can propagate backwards towards the source. This phenomenon is known as congestion spreading and an example is shown in the following sequence of diagrams.

Figure 1 is an example of a SAN that is not experiencing congestion. Both Host 1 and Host 2 are performing READ commands to the array.
Since both the array and host are attached at 16Gbps and there is sufficient ISL bandwidth (i.e., 32G), there is no congestion in the SAN.

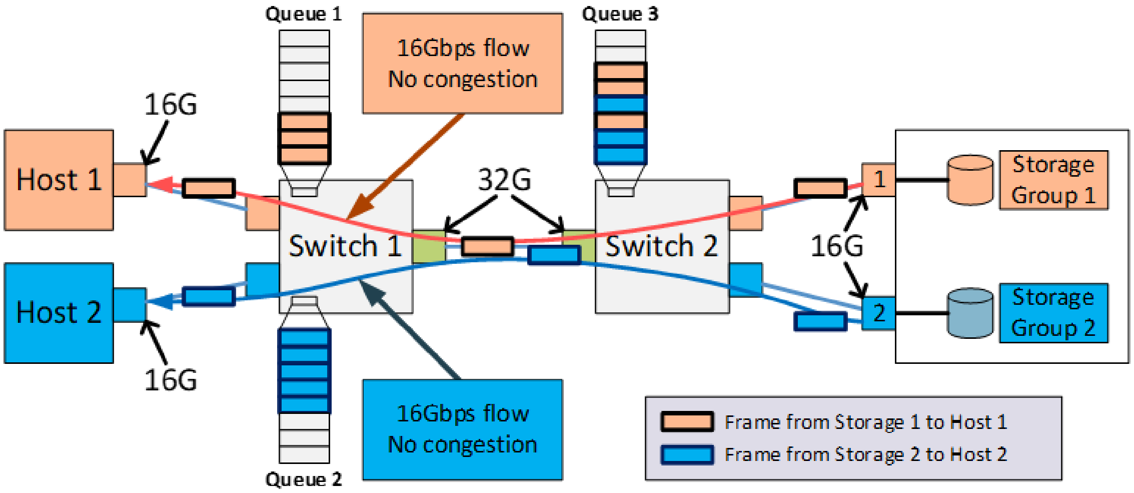


Figure 1 No Congestion

Figure 2 shows an example of a SAN that is experiencing Congestion Spreading due to Oversubscription. Note that the only difference between the two figures is that in Figure 3, the interface on Host 1 was set to run at 4Gbps instead of 16Gbps. As soon as this is done, if the array interface transmits data at a rate that is anything greater than the speed of the attached HBA (i.e., 4G), Host 1 will be unable to receive the data at the rate that's being transmitted, and the immediate impact is the queuing of frames. As Queue 1 fills, the congestion spreads back to the source of the data. Since both Host 1 and Host 2 are sharing the same Inter Switch Link (ISL), this congestion impacts the "innocent flow" between Host 2 and Storage 2, reducing throughput from 16Gbps to 4Gbps.

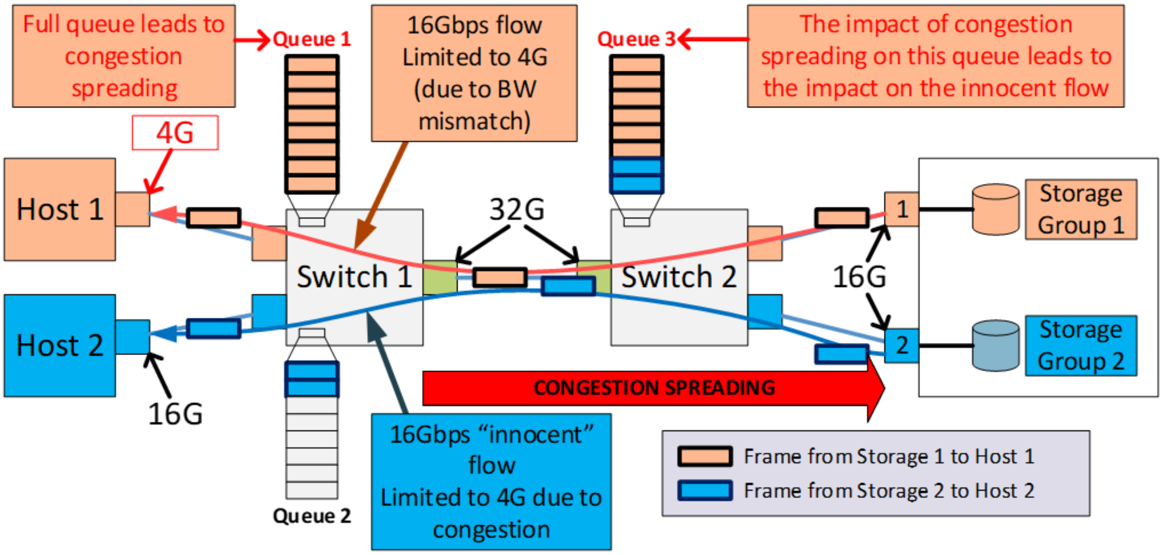

Full queue leads to congestion spreading
Queue 1
16Gbps flow Limited to 4G (due to BW mismatch)
Queue 3
The impact of congestion spreading on this queue leads to the impact on the innocent flow
4G
Host 1
32G
Switch 1
Switch 2
1
Storage Group 1
16G
Host 2
2
Storage Group 2
16G
CONGESTION SPREADING
16Gbps "innocent" flow Limited to 4G due to congestion
Frame from Storage 1 to Host 1
Frame from Storage 2 to Host 2
Queue 2

Figure 2 Congestion

Additional information about congestion and congestion spreading can be found in the Congestion and Backpressure section of the *Networked Storage Concepts and Protocols* Techbook: (https://www.emc.com/collateral/hardware/technical-documentation/h4331-networked-storage-cncpts-prtcls-sol-gde.pdf). It's important to note that oversubscription is only one of the potential causes of congestion spreading. Other causes will be explained in the following sections.

- Congestion Ratio (c ratio)

The Congestion Ratio or c ratio is a calculated value that can help detect when congestion spreading is occurring. For example, Figure 3 depicts a host (i.e., Host 1) that is capable of receiving data at a rate of 4Gbps but is receiving data from a storage interface that is capable of transmitting data at 16Gbps.

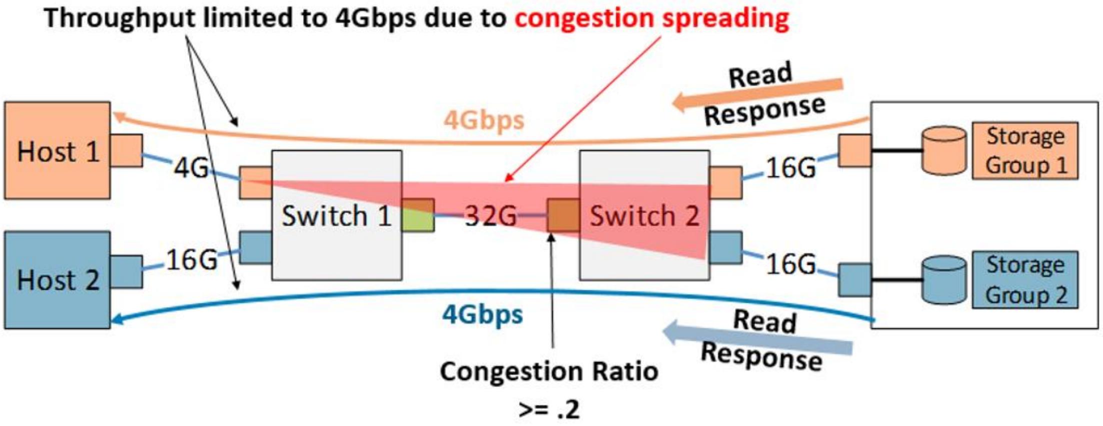


Figure 3 Congestion Ratio

Part of what makes these problems so difficult to detect and troubleshoot is that from the perspective of the 4G interface on switch 1, everything is fine. The switch interface is transmitting frames as fast as the link allows. However, since the storage is transmitting data at the rate it's link allows (i.e., 16Gbps), there is going to be 12Gbps (16Gbps - 4Gbps) of bandwidth that will be transmitted by the array and will need to be queued somewhere. This queuing typically happens in the Fabric and is the cause of Congestion Spreading. As mentioned above, one method that can be used to detect the presence of Congestion Spreading is to calculate the congestion ratio. To do this, take the "Time Spent at zero transmit credit" counter and divide it by the Frames Transmitted counter and you will have a number (typically between 0 and 1). If this number is greater than .2, you have congestion. By the way, this number needs to be calculated on a per interface basis, so it's probably best just to script the process for checking this value.

4 Congestion Spreading Due to Oversubscription

The following case study is based on congestion spreading due to oversubscription. The topology for this case study is shown in Figure 4 below. In this case study, you will learn about the tools and techniques that are currently available to help detect and prevent this issue from occurring.

> **Note:** Congestion Spreading is an extremely difficult problem to detect and resolve. This is primarily due to the inability of the current generation of management tools to provide a clear indication that the issue is occurring, let alone providing any guidance on how to actually solve the problem. As a result, troubleshooting these problems requires the end user to understand what the problem is, and then know how to use the tools that are currently available to draw conclusions from the limited data available.

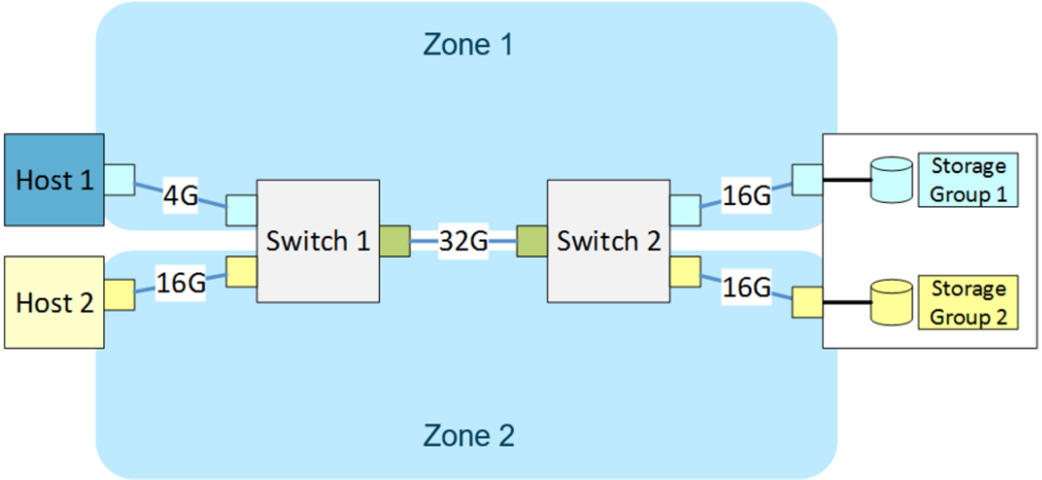


Figure 4 Topology of Oversubscription Case Study

- Scenario:

User 1 has had an existing application running on Host 2 (16G HBA) that is running I/O at various block sizes, queue depths and I/O patterns. This application has been running for a long time in this environment and has not experienced any issues until recently. Earlier this month, User 2 decided to load an application on Host 1 (4G HBA) for testing. Initially everything was fine in the environment with regards to performance and latency. However, User 1 recently started to notice performance issues with their application.

- Troubleshooting Overview:

To troubleshoot any problem in general, you must first understand how things perform and are configured when working in their ideal conditions. As you know, a SAN has many moving parts that make up the ecosystem, so it's very important to build out an environment profile consisting of profiles for the 3 major components that make up a SAN: Application(s), SAN Fabric and Storage.

Building these baseline profiles at various components in your environment will give you the necessary ability to easily pinpoint issues when they arise. It should be noted that these profiles are not a one-time thing. You should constantly gather baseline line data throughout the lifetime of your environment so that you not only troubleshoot issues, but plan for further grow and expansion.

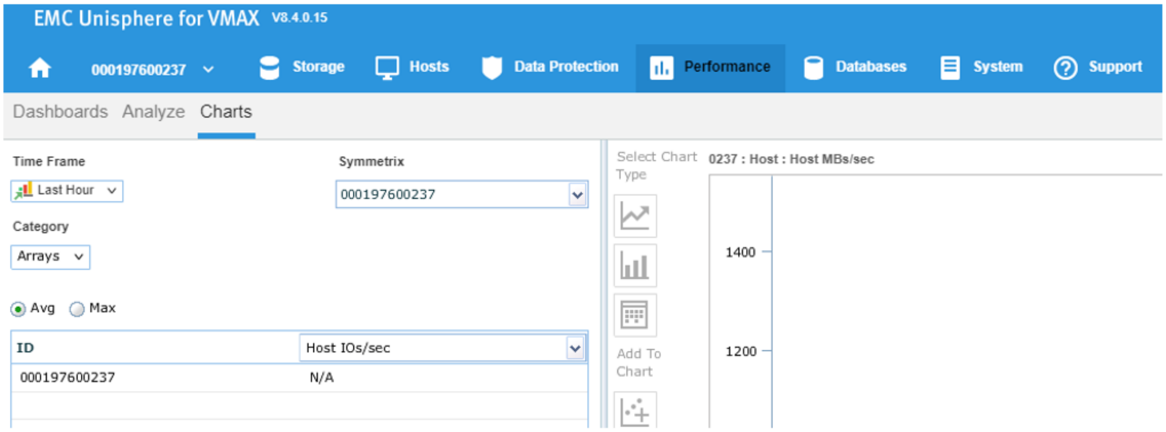
In the next few sections, we will show that you must gather these baseline statistics from your storage array, so that when an issue does arise as stated in the scenario above, you are well equipped to find the root cause.

Application Base Line

With Dell EMC VMAX/PowerMax, when you enable Performance monitoring you can go back into your history (up to a year since enabling the feature) so that you can understand what your application profile was in terms of IOPs and response times before any changes were made. Having this application base profile will allow you to use the charts generated and easily determine where there might be issues.

Generating Application Base Profile Graphs

1. In Unisphere, click on **Performance >Charts**



Time Frame: Last Hour

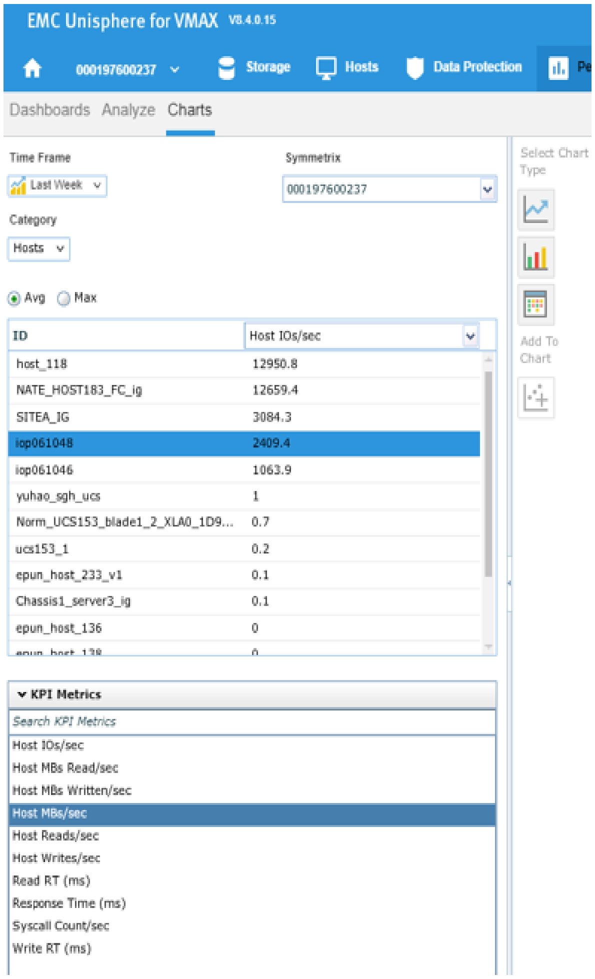
Symmetrix: 000197600237

Category: Arrays

● Avg ○ Max

ID	Host IOs/sec
000197600237	N/A

2. Select a **Time Frame**. This would be any time BEFORE you noticed the performance issue. Under the Category drop down menu, select **Hosts > Hosts.**



Time Frame: Last Week

Symmetrix: 000197600237

Category: Hosts

● Avg ○ Max

ID	Host IOs/sec
host_118	12950.8
NATE_HOST183_FC_ig	12659.4
SITEA_IG	3084.3
iop061048	2409.4
iop061046	1063.9
yuhao_sgh_ucs	1
Norm_UCS153_blade1_2_XLA0_1D9...	0.7
ucs153_1	0.2
epun_host_233_v1	0.1
Chassis1_server3_ig	0.1
epun_host_136	0
epun_host_138	0

KPI Metrics

Search KPI Metrics

Host IOs/sec
Host MBs Read/sec
Host MBs Written/sec
Host MBs/sec
Host Reads/sec
Host Writes/sec
Read RT (ms)
Response Time (ms)
Syscall Count/sec
Write RT (ms)

3. Select the Host in question. For **KPI Metrics** we will generate seven (7) different charts. Repeat this section for each KPI metric. If you click all the metrics at once, the chart will place them in a single graph.
   a. Host IOs/sec
   b. Host MBs/sec
   c. Host Reads/sec
   d. Host Writes/sec
   e. Read RT (ms)
   f. Response Time (ms)
   g. Write RT (ms)

In Figure 5, “Host IOs and MBs/secs,”, we are looking at the Host IOs and MBs/sec. From these charts we can see when and for how long the application has been running the most IOs and utilizing the full bandwidth of the link as well as the low points.

Note: In the legend you will notice there are two hosts but currently we are only showing the IO for one of the hosts because the other host is not doing any IO"

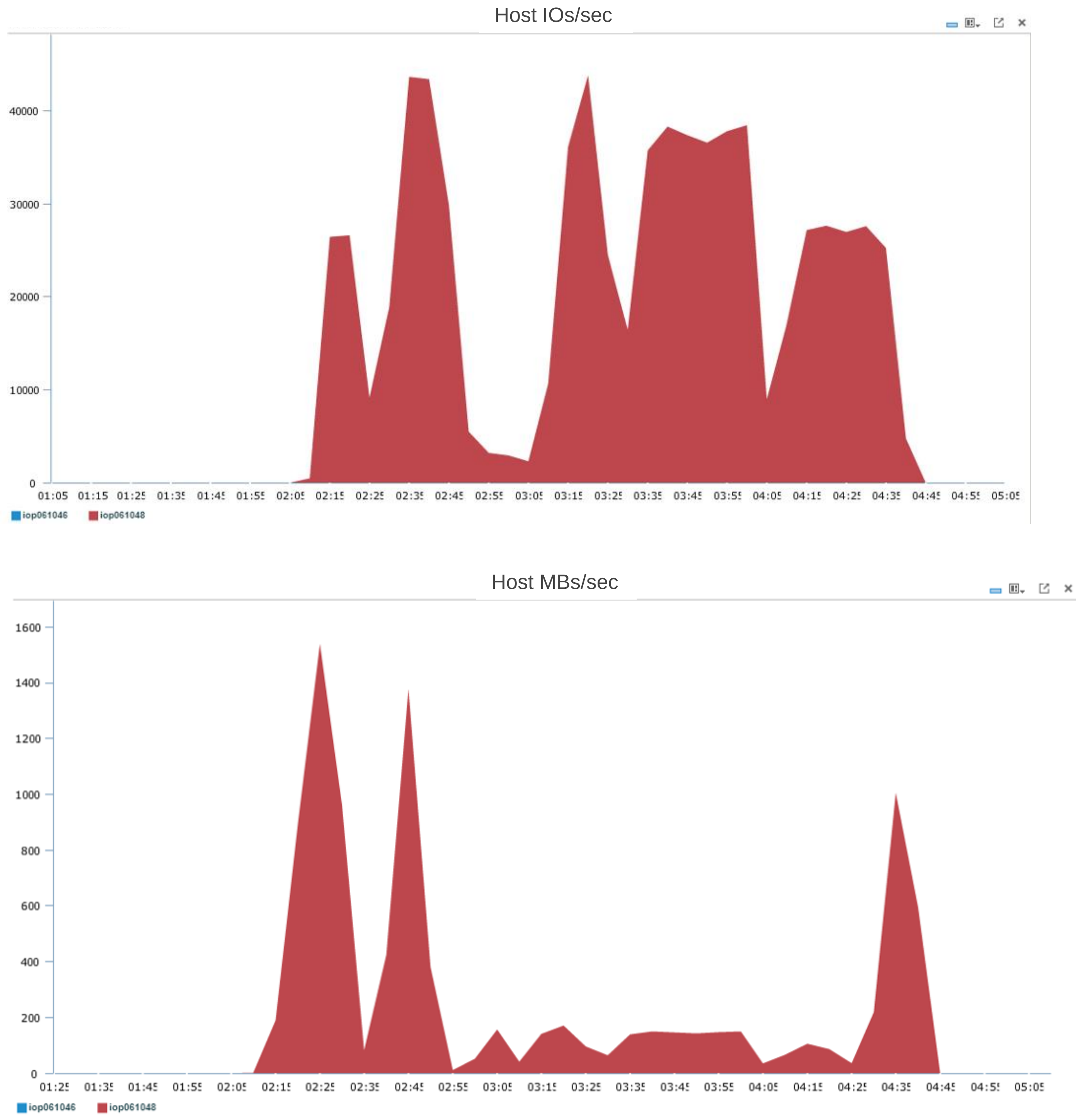


Figure 5 Host IOs and MBs/secs

The charts shown in Figure 6, “Read and Writes/secs,” provide a breakdown of the type of IO the application generates. Based on these charts, we can determine what percentage of the application IOs are READs versus WRITEs. In this case we can confirm that the application is about 70/30 in terms of READs/WRITEs.

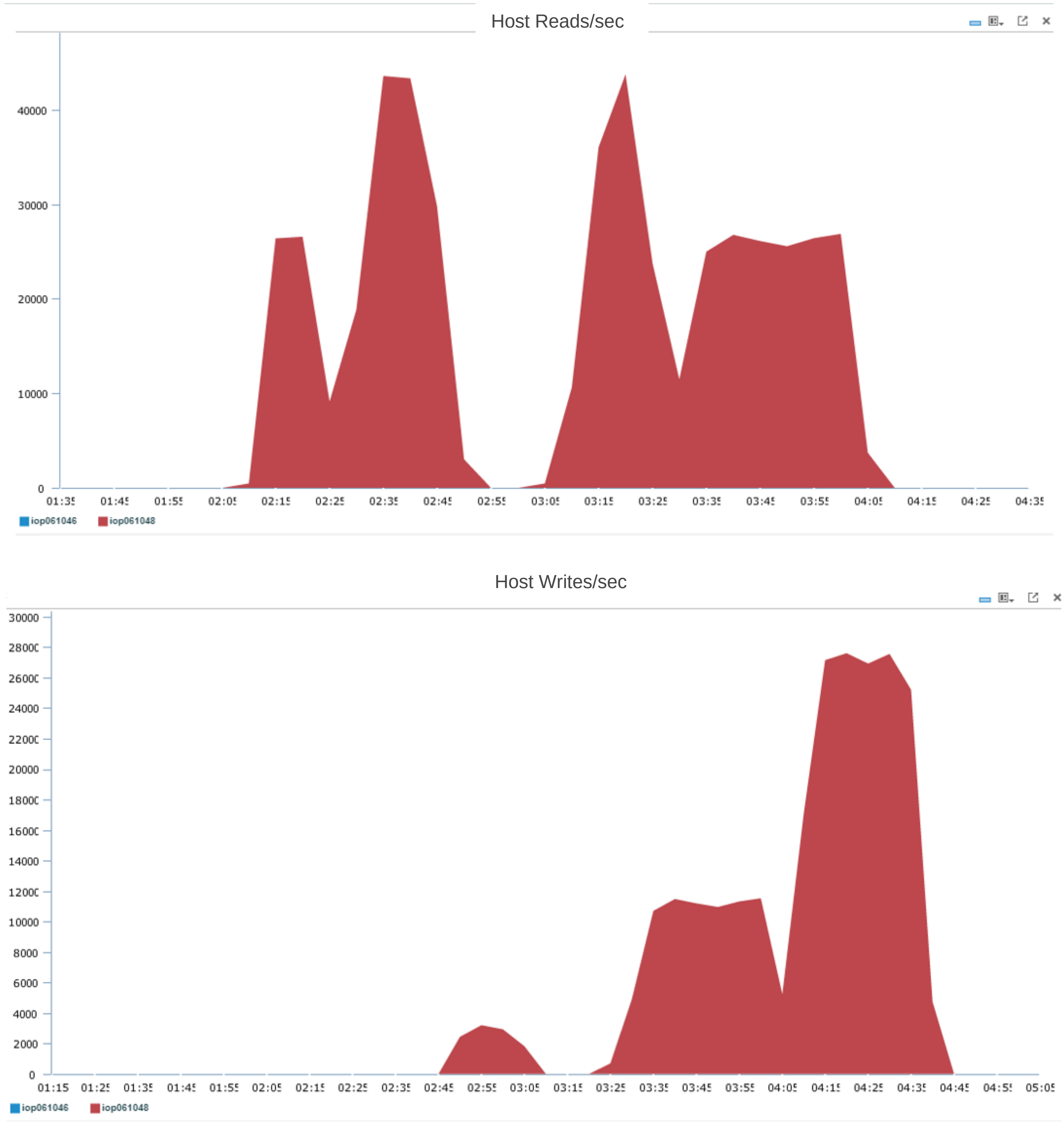


Figure 6 Read and Writes/secs

Figure 7 and Figure 8 are probably the most useful charts to use when troubleshooting. They provide a breakdown of response times between READs and WRITEs that allows us to understand the latency that the application is experiencing. This is extremely useful for when we need to troubleshoot performance issues, because if there is a spike in response times, we can correlate the spike back to specific events using the previous charts.

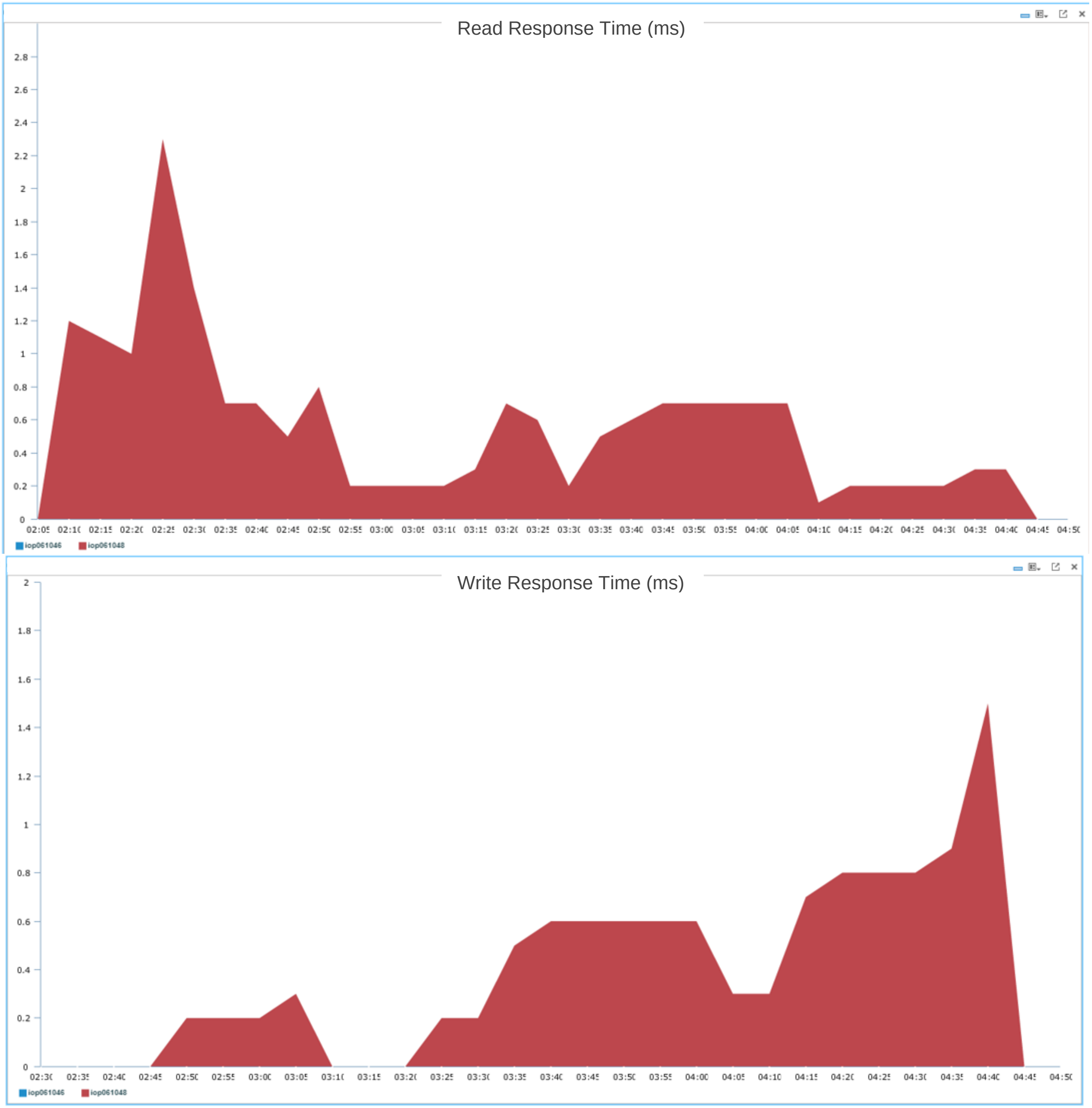


Figure 7 Read and Write RT (ms)

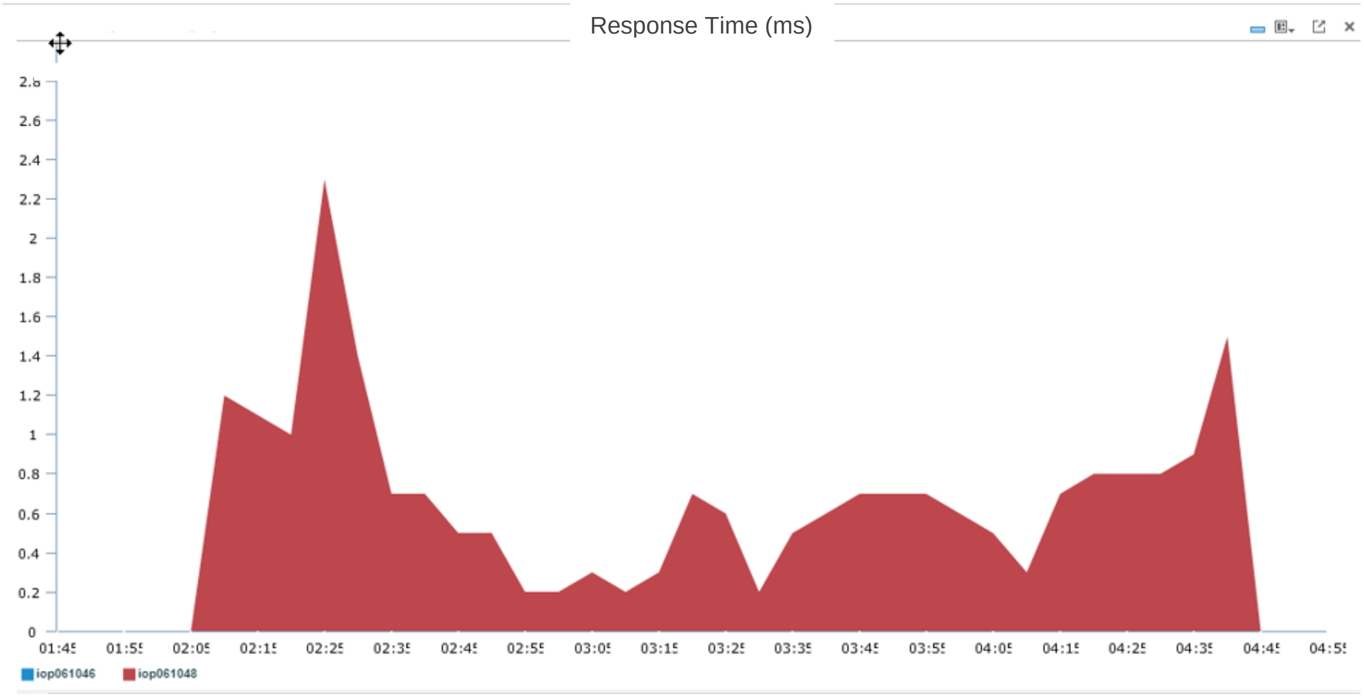


Figure 8 Response Times (ms)

Now we have an application profile for User 1's app; we know it's about 70/30 Reads/Writes with response times on average of ~0.7ms with a max of 2.3ms response time.

As discussed in the scenario section, we recently added a new host that started a performance issue in the environment, so let’s look at how we can troubleshoot that issue.

Since we are having a performance issue in our environment, we will need to implement the features available in the SAN that can help us determine when these types of issues arise.

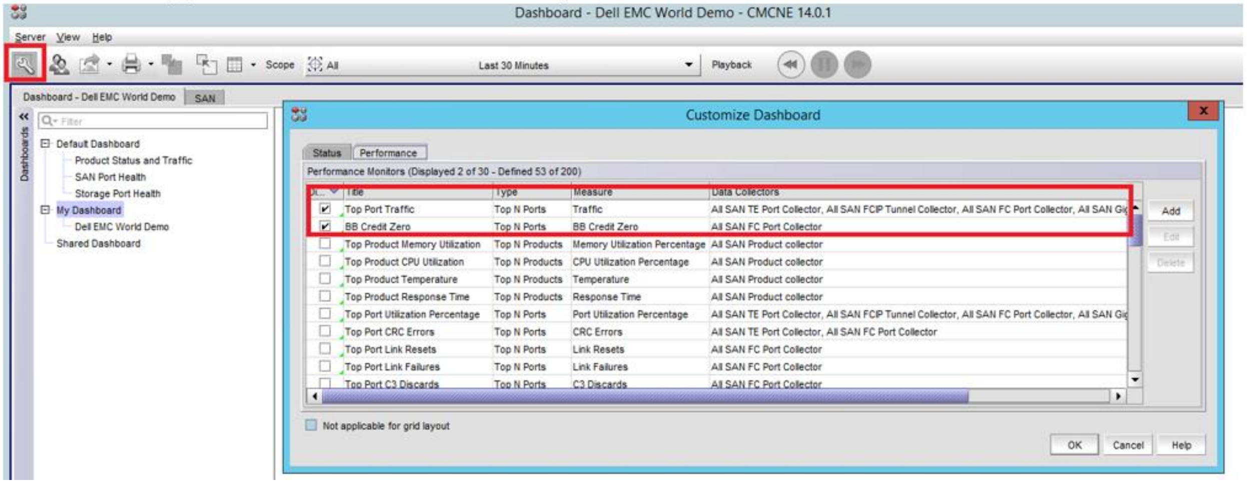
Because we know what our average response times are, based on the application profile, we know that this performance issue is higher than the expected response times.

Connectrix SAN Congestion Spreading Alerts

In this section, we will review the type of congestion events that we reported from the SAN switch side. Please ensure that you have completed enabling these features in the environment per the Prerequisites.

4.1.1 Brocade

1. Ensure that you have at least **Top Port Traffic** and **BB Credit Zero** on your dashboard. You can click the wrench in the upper left-hand corner to add them if you do not.



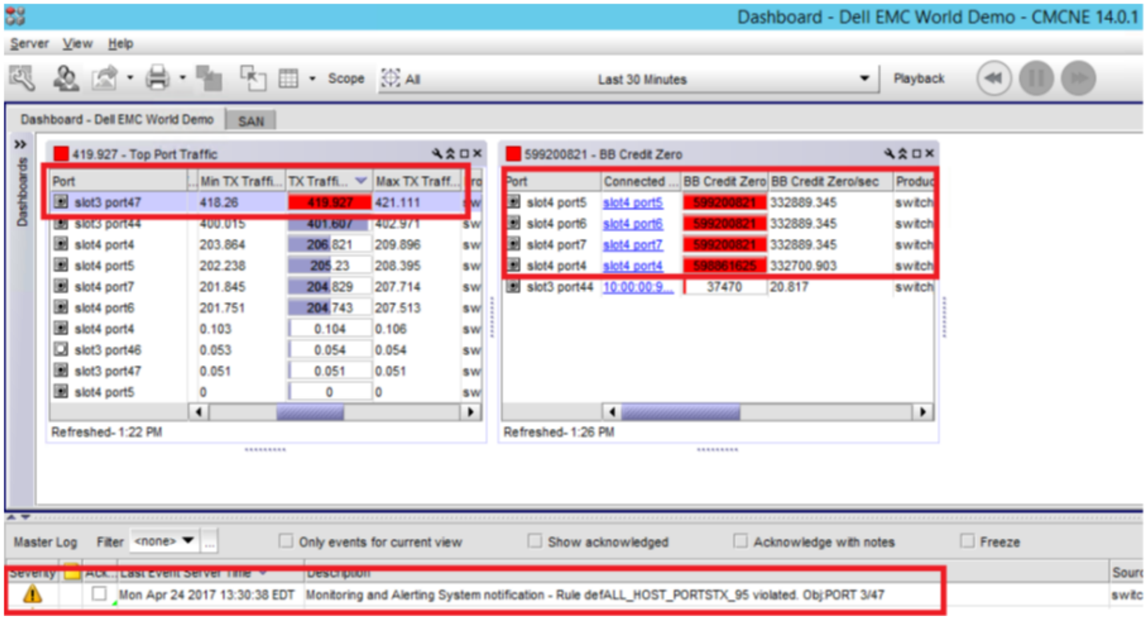
Performance Monitors (Displayed 2 of 30 - Defined 53 of 200)

Di...	Title	Type	Measure	Data Collectors
☑	Top Port Traffic	Top N Ports	Traffic	All SAN TE Port Collector, All SAN FCIP Tunnel Collector, All SAN FC Port Collector, All SAN Gi
☑	BB Credit Zero	Top N Ports	BB Credit Zero	All SAN FC Port Collector
☐	Top Product Memory Utilization	Top N Products	Memory Utilization Percentage	All SAN Product collector
☐	Top Product CPU Utilization	Top N Products	CPU Utilization Percentage	All SAN Product collector
☐	Top Product Temperature	Top N Products	Temperature	All SAN Product collector
☐	Top Product Response Time	Top N Products	Response Time	All SAN Product collector
☐	Top Port Utilization Percentage	Top N Ports	Port Utilization Percentage	All SAN TE Port Collector, All SAN FCIP Tunnel Collector, All SAN FC Port Collector, All SAN Gi
☐	Top Port CRC Errors	Top N Ports	CRC Errors	All SAN TE Port Collector, All SAN FC Port Collector
☐	Top Port Link Resets	Top N Ports	Link Resets	All SAN FC Port Collector
☐	Top Port Link Failures	Top N Ports	Link Failures	All SAN FC Port Collector
☐	Top Port C3 Discards	Top N Ports	C3 Discards	All SAN FC Port Collector

☐ Not applicable for grid layout

Figure 9 CMCNE Dashboard

2. When congestion spreading due to oversubscription occurs, as in the example in Figure 9, "CMCNE Dashboard", you will typically see the following alerts in CMCNE dashboard:
   a. Highly utilized F-Port
   b. BB Credit Zero


Dashboard - Dell EMC World Demo - CMCNE 14.0.1
Server View Help
Scope All Last 30 Minutes Playback
Dashboard - Dell EMC World Demo SAN
Dashboards
Master Log Filter <none>
☐ Only events for current view
☐ Show acknowledged
☐ Acknowledge with notes
☐ Freeze

419.927 - Top Port Traffic

Port	Min TX Traffi...	TX Traffi...	Max TX Traff...	ro
slot3 port47	418.26	419.927	421.111	sw
slot3 port44	400.015	401.607	402.971	sw
slot4 port4	203.864	206.821	209.896	sw
slot4 port5	202.238	205.23	208.395	sw
slot4 port7	201.845	204.829	207.714	sw
slot4 port6	201.751	204.743	207.513	sw
slot4 port4	0.103	0.104	0.106	sw
slot3 port46	0.053	0.054	0.054	sw
slot3 port47	0.051	0.051	0.051	sw
slot4 port5	0	0	0	sw

Refreshed- 1:22 PM

599200821 - BB Credit Zero

Port	Connected ...	BB Credit Zero	BB Credit Zero/sec	Produc
slot4 port5	slot4 port5	599200821	332889.345	switch
slot4 port6	slot4 port6	599200821	332889.345	switch
slot4 port7	slot4 port7	599200821	332889.345	switch
slot4 port4	slot4 port4	598861625	332700.903	switch
slot3 port44	10:00:00:9...	37470	20.817	switch

Refreshed- 1:26 PM

Severity		Ack...	Last Event Server Time	Description	Sourc
⚠		☐	Mon Apr 24 2017 13:30:38 EDT	Monitoring and Alerting System notification - Rule defALL_HOST_PORTSTX_95 violated. Obj:PORT 3/47	switc

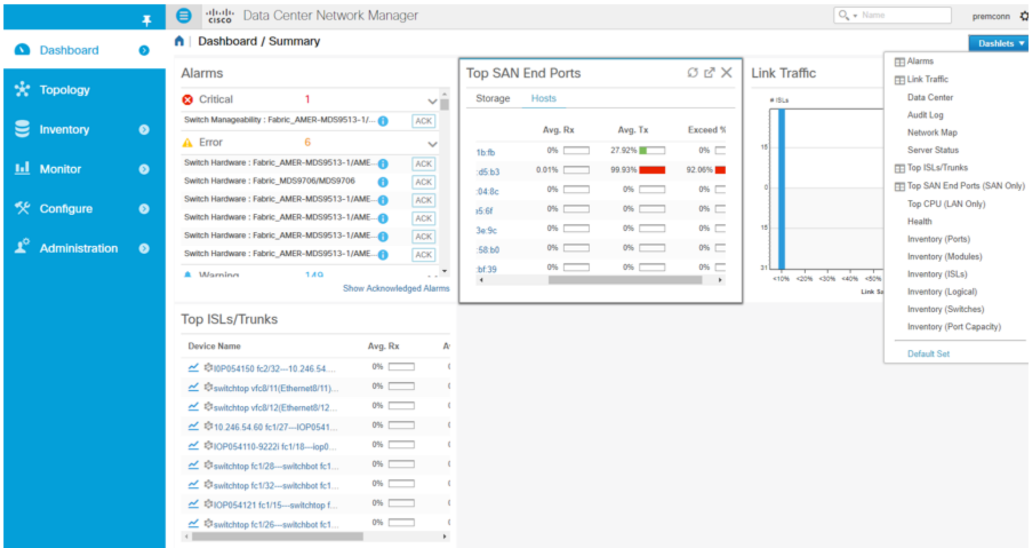
Figure 10 CMCNE Dashboard displaying alerts

3. The combination of these two events—highly utilized F-port and high buffer-to-buffer credits going zero on the ISLs can indicate you have a potential performance issue that needs to be investigated. Refer to the Remediation chapter for steps on what to review.

4.1.2 Cisco

1. In the DCNM dashboard. You should have **Top SAN End Ports** as a dashlet. If not, you can add it from the drop-down menu. In the **Top SAN End Ports,** you will see a device(s) reporting to be over 90% utilized. DCNM will have default thresholds that cause it to flag a port either yellow or red when it starts to exceed the default utilization. This alert by itself does not necessarily mean that there is a performance issue in the SAN. We will need to look for other alerts in the fabrics as well.

–



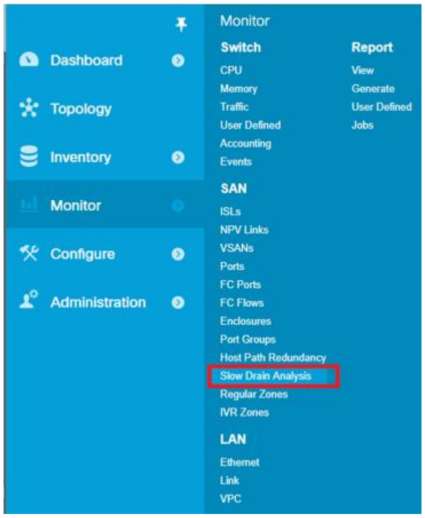
Switch Manageability : Fabric_AMER-MDS9513-1/...
Switch Hardware : Fabric_AMER-MDS9513-1/AME...
Switch Hardware : Fabric_MDS9706/MDS9706
Switch Hardware : Fabric_AMER-MDS9513-1/AME...
Switch Hardware : Fabric_AMER-MDS9513-1/AME...
Switch Hardware : Fabric_AMER-MDS9513-1/AME...
Switch Hardware : Fabric_AMER-MDS9513-1/AME...

	Avg. Rx	Avg. Tx	Exceed %
1b:fb	0%	27.92%	0%
:d5:b3	0.01%	99.93%	92.06%
:04:8c	0%	0%	0%
i5:6f	0%	0%	0%
3e:9c	0%	0%	0%
:58:b0	0%	0%	0%
:bf:39	0%	0%	0%

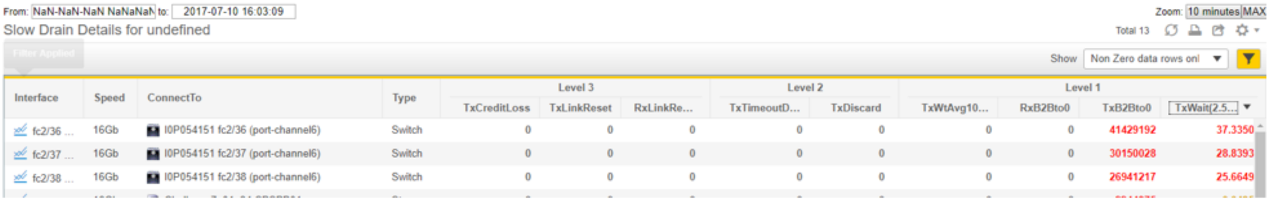
Device Name	Avg. Rx
I0P054150 fc2/32---10.246.54....	0%
switchtop vfc8/11(Ethernet8/11)...	0%
switchtop vfc8/12(Ethernet8/12...	0%
10.246.54.60 fc1/27---IOP0541...	0%
IOP054110-9222i fc1/18---iop0...	0%
switchtop fc1/28---switchbot fc1...	0%
switchtop fc1/32---switchbot fc1...	0%
IOP054121 fc1/15---switchtop f...	0%
switchtop fc1/26---switchbot fc1...	0%

–

2. If you see a highly utilized F port, run the Slow Drain Analysis tool.
 Click on **Monitor—>SAN > Slow Drain Analysis.**



3. Run the slow drain analysis tool for 10 minutes. When the report finishes you will notice that there is a large amount of TxWait counter incrementing during the time when the report was running. The combination of these alerts and the highly utilized F-port indicates that there is SAN congestion occurring due to oversubscription.



From: NaN-NaN-NaN NaNaNaN to: 2017-07-10 16:03:09

Zoom: 10 minutes MAX

Slow Drain Details for undefined

Total 13

Show Non Zero data rows onl

Interface	Speed	ConnectTo	Type	Level 3			Level 2		Level 1			
				TxCreditLoss	TxLinkReset	RxLinkRe...	TxTimeoutD...	TxDiscard	TxWtAvg10...	RxB2Bto0	TxB2Bto0	TxWait(2.5...)
fc2/36 ...	16Gb	I0P054151 fc2/36 (port-channel6)	Switch	0	0	0	0	0	0	0	41429192	37.3350
fc2/37 ...	16Gb	I0P054151 fc2/37 (port-channel6)	Switch	0	0	0	0	0	0	0	30150028	28.8393
fc2/38 ...	16Gb	I0P054151 fc2/38 (port-channel6)	Switch	0	0	0	0	0	0	0	26941217	25.6649

The combination of these two events – highly utilized F-port and high buffer-to-buffer credits going zero on the ISLs – can indicate that you have a potential performance issue that should be investigated. Refer to the Remediation section for steps on what to review.

VMAX UNISPHERE CONGESTION SPREADING ALERTS

In this section, we will learn how to use the Unisphere for VMAX/PowerMAX to correlate the SAN switch events to the storage array.

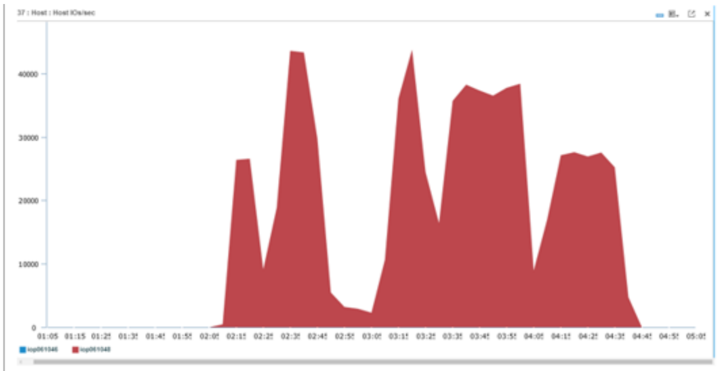
Please make sure that you have completed enabling of these features in the environment per the Prerequisites section.

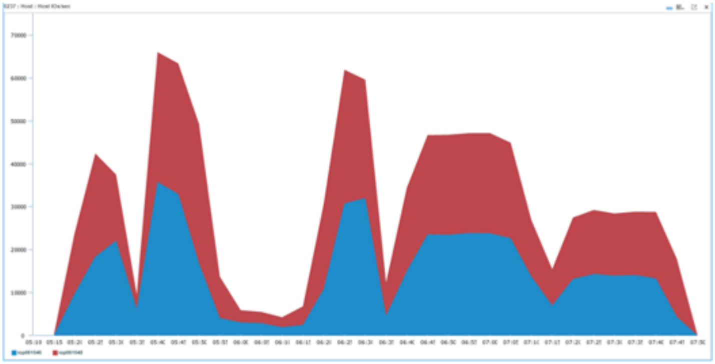
1. Use the steps from Generating Application Base Profile Graphs. Generate the same seven (7) charts, adding User 2's host into the mix (because that was one of the recent changes in the environment before the performance issue occurred). Review the data.

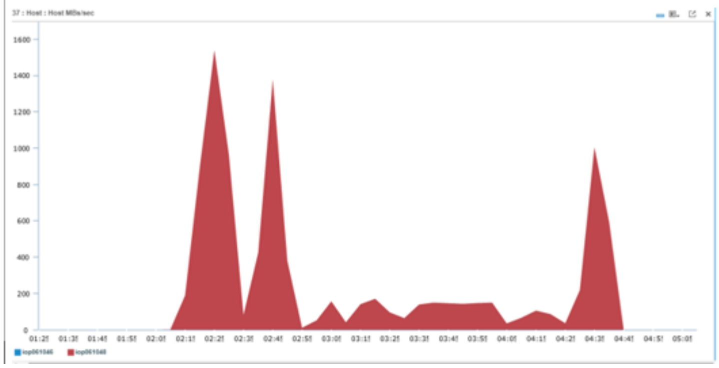
Keep in mind that our Application Base Profile = 70/30 Reads/Writes with average response times of ~0.7ms with a max of 2.3ms response time.

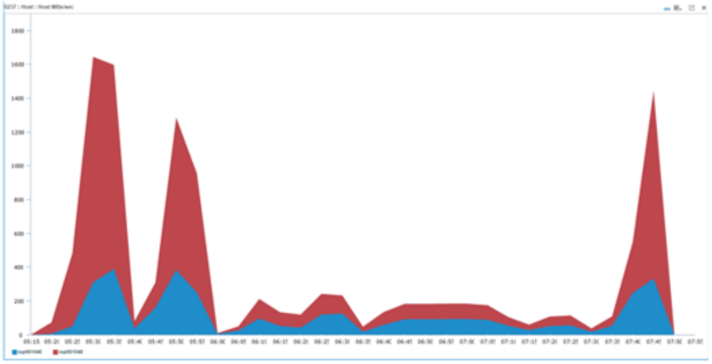
In Figure 11 below, when we compare the IOs and MBs/sec, we don't really see an indication of an issue. In fact, if you compare it to the original base line application chart, we are in fact doing more IOPs.

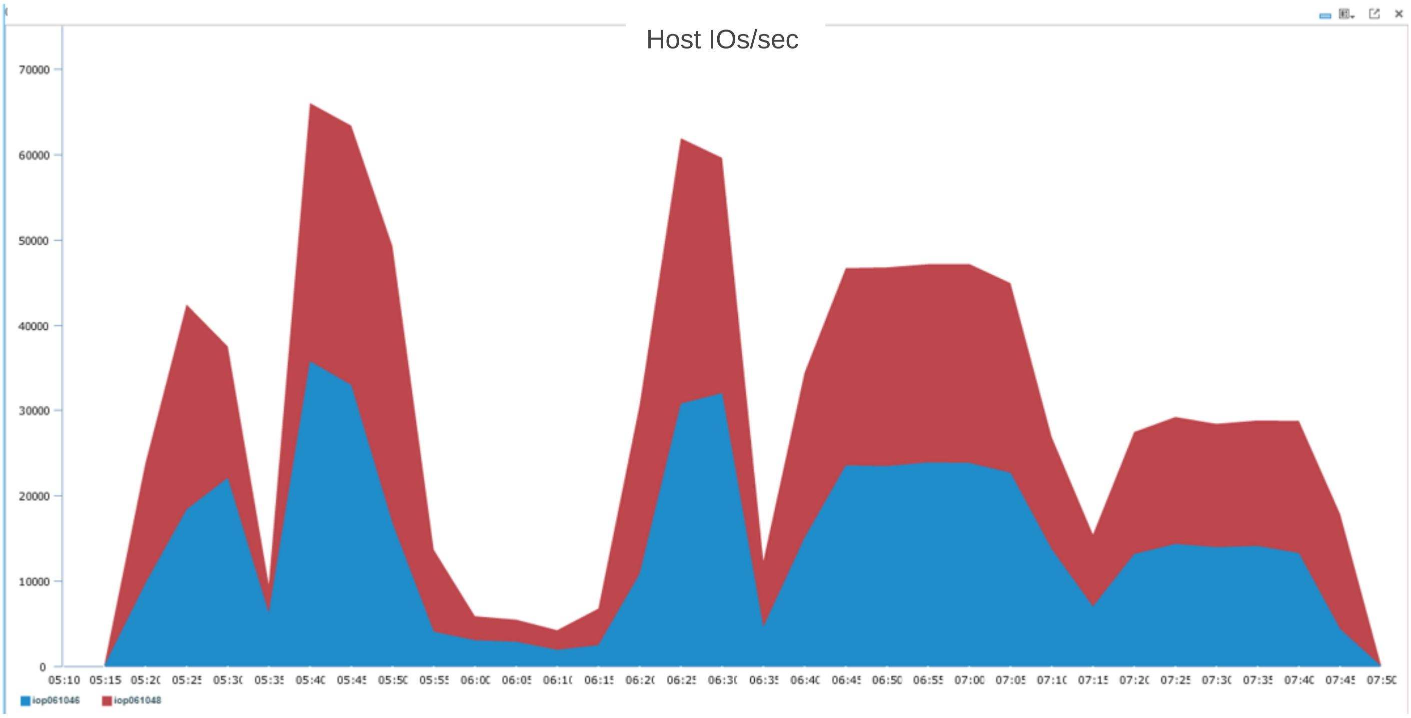
In addition, you can see there are some points where we are getting close to line rate (highlighted below). These points will come into play later.


37 : Host : Host IOs/sec
40000
30000
20000
10000
0
01:05 01:15 01:2! 01:3! 01:4! 01:5! 02:0! 02:1! 02:2! 02:3! 02:4! 02:5! 03:0! 03:1! 03:2! 03:3! 03:4! 03:5! 04:0! 04:1! 04:2! 04:3! 04:4! 04:5! 05:0!
iop061046
iop061048


0237 : Host : Host IOs/sec
70000
60000
50000
40000
30000
20000
10000
0
05:10 05:15 05:2(05:2! 05:3(05:3! 05:4(05:4! 05:5(05:5! 06:0(06:0! 06:1(06:1! 06:2(06:2! 06:3(06:3! 06:4(06:4! 06:5(06:5! 07:0(07:0! 07:1(07:1! 07:2(07:2! 07:3(07:3! 07:4(07:4! 07:5(
iop061046
iop061048


37 : Host : Host MBs/sec
1600
1400
1200
1000
800
600
400
200
0
01:2! 01:3! 01:4! 01:5! 02:0! 02:1! 02:2! 02:3! 02:4! 02:5! 03:0! 03:1! 03:2! 03:3! 03:4! 03:5! 04:0! 04:1! 04:2! 04:3! 04:4! 04:5! 05:0!
iop061046
iop061048


0237 : Host : Host MBs/sec
1800
1600
1400
1200
1000
800
600
400
200
0
iop061046
iop061048


Host IOs/sec
70000
60000
50000
40000
30000
20000
10000
0
05:10 05:15 05:2(05:2! 05:3(05:3! 05:4(05:4! 05:5(05:5! 06:0(06:0! 06:1(06:1! 06:2(06:2! 06:3(06:3! 06:4(06:4! 06:5(06:5! 07:0(07:0! 07:1(07:1! 07:2(07:2! 07:3(07:3! 07:4(07:4! 07:5(
iop061046
iop061048

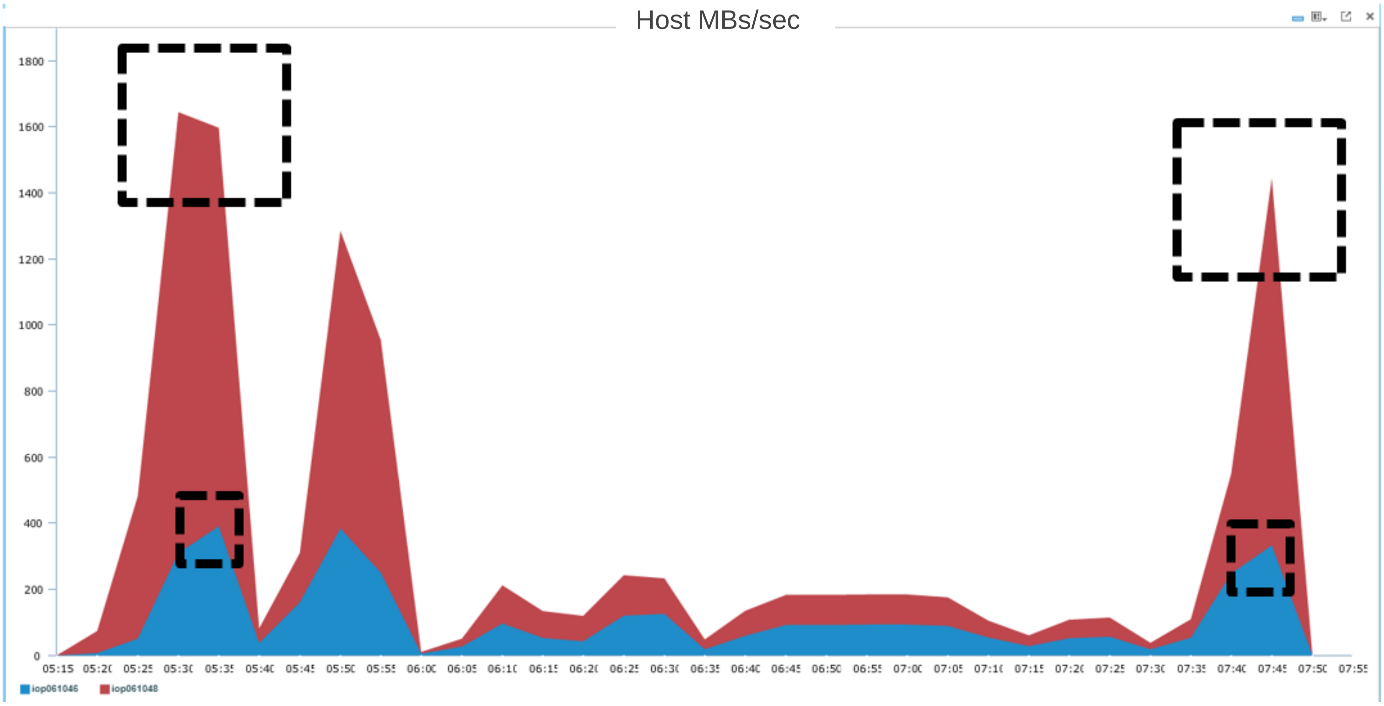


Figure 11 Host IOs and MB/secs

Figure 12 shows the compare in Reads/Writes between the two servers as you can see there is not much difference between the IO profile between the servers at this point.

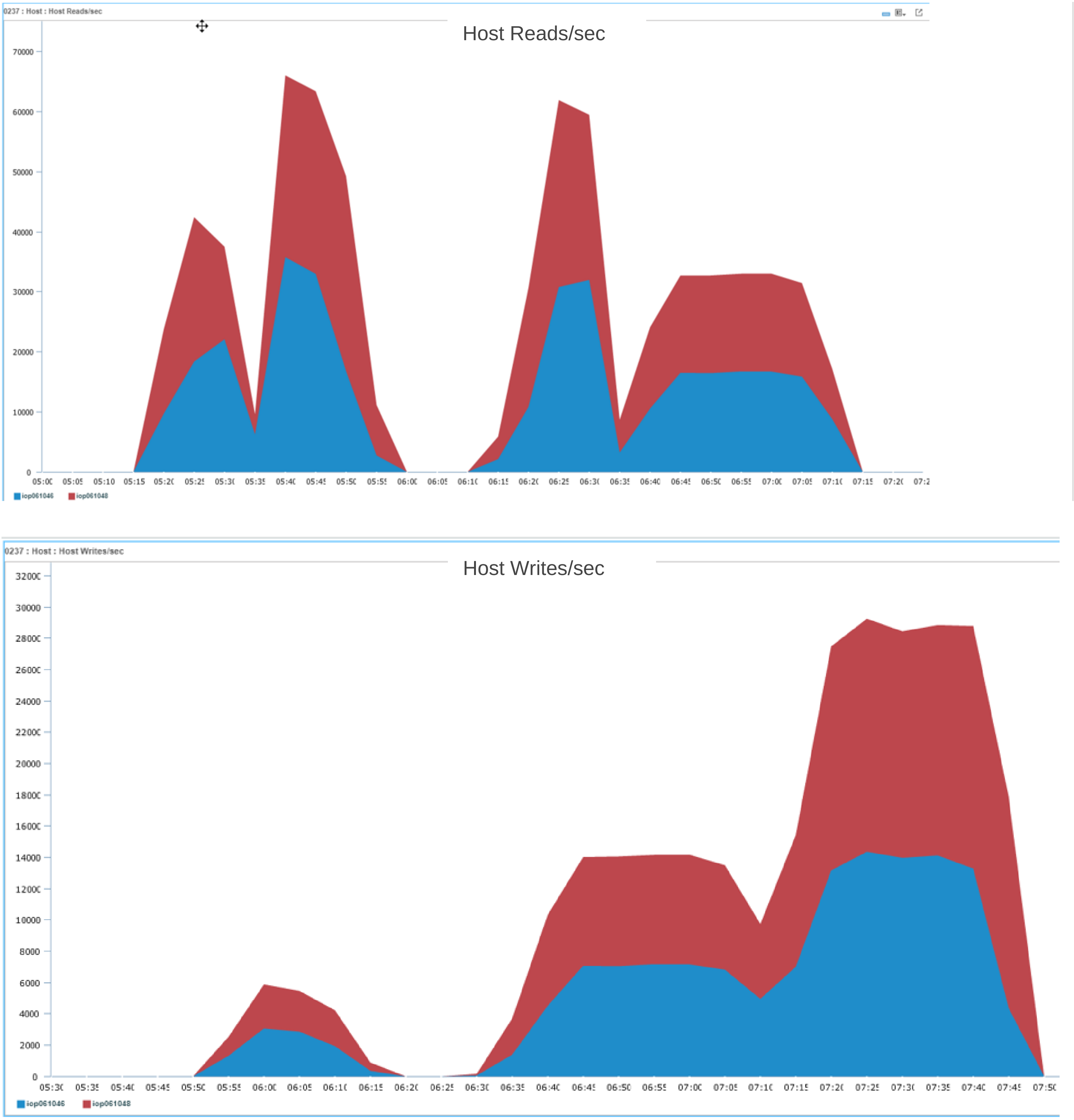
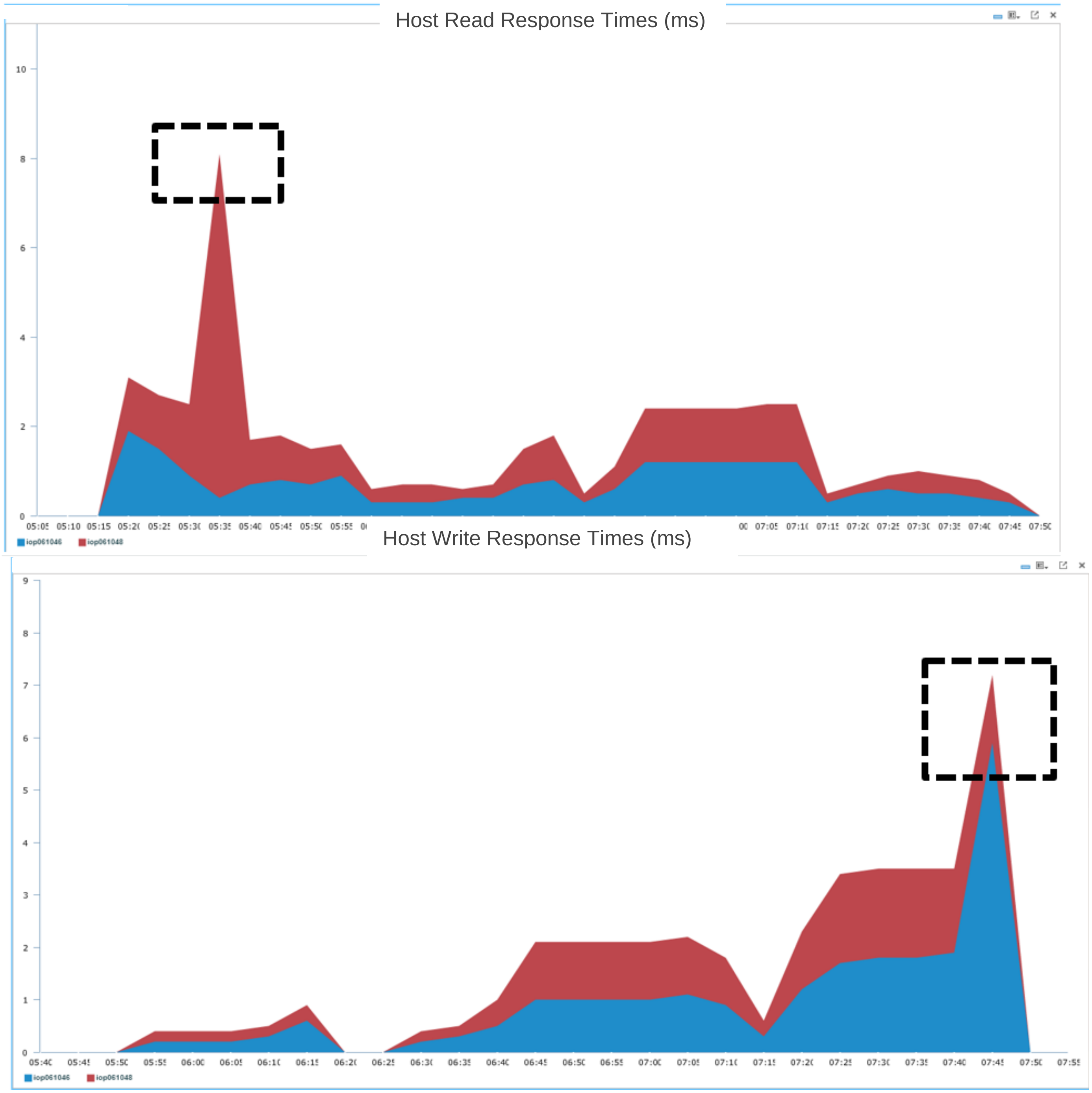


Figure 12 Host Reads and Writes/secs

Figure 13 provides the most useful information. Keeping our application profile in mind, we are seeing Response times of average of ~0.7ms and max of 2.3ms. From the chart below, we can see that there is a massive spike in response times where we get into the 8ms range, and our overall average response times have increased as well.

Looking back at Figure 11, we can see that these high response times correlate back to when the both servers are close to line rate.
Typically, in Fibre Channel, you will need large block IOs (greater than 128k) to saturate a link.


Host Read Response Times (ms)

Host Write Response Times (ms)

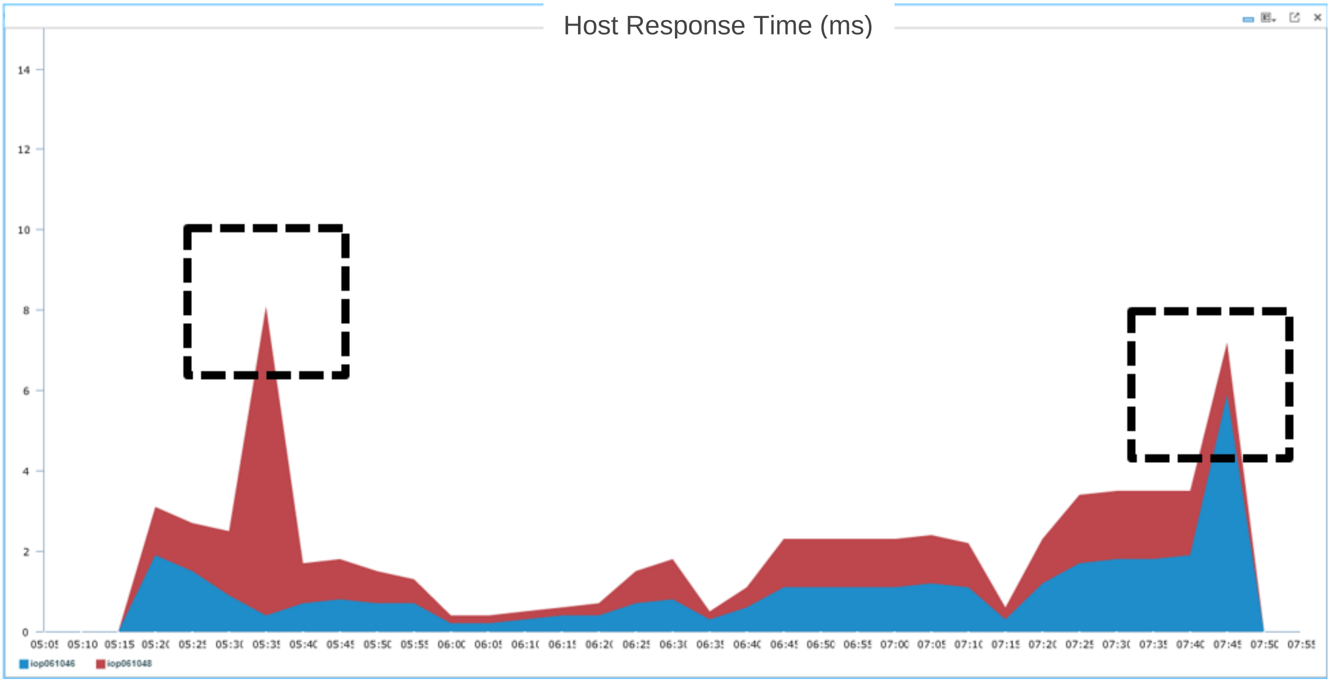


Figure 13 Host Read and Write Response Times (ms)

CONCLUSION

To re-cap all the information we know thus far in this case study:

- Connectrix SAN:

1. Our SAN is reporting high numbers of buffer-to-buffer credits going to zero.
2. We are seeing high traffic utilization on our F-port(s).

- Dell EMC VMAX/PowerMAX:

1. High response times during full link utilization

As stated earlier, congestion due to bandwidth mismatch is extremely difficult to detect and confirm with the set of tools available today. However, based on the above alerts, we can infer the issue is due to bandwidth mismatch and large block Reads/Write. This is indicated by the high response times during full link utilization.

Another way to detect this issue is by using the congestion ratio. Today we must calculate this manually in the environment (or you can attempt to script it) but we know once your C-ratio is greater than .2, you will experience congestion because of the backpressure that is occurring in the SAN environment. The C-ratio would be your first indication of a slow drain.

5 Remediation

PREVENTION

For this specific case study (Congestion Spreading due to Oversubscription) there are a few options you can deploy in your environment to help prevent this issue from occurring.

Bandwidth Ratio

When reviewing the SAN, you want to identify devices that are running at lower speeds and then understand their type of application traffic profiles. Remember, just because you have a bandwidth mismatch, does NOT mean there is necessarily an issue.
Review the fabric end-to-end to ensure all end devices are running at the same link speeds.
Ensure ample amount of bandwidth on your ISLs. A good rule of thumb is the total ISL bandwidth should be equal to or greater than the total amount of storage bandwidth in the fabric where possible.
You can modernize your entire SAN by ensuring that you upgrade all components end-to-end as shown in Figure 14 below. The con with this approach is that zero oversubscription from end-to-end is impractical in larger environments. In addition, it can be very expensive. Therefore, you should just focus on upgrading the specific host, switch and storage connectivity as well.

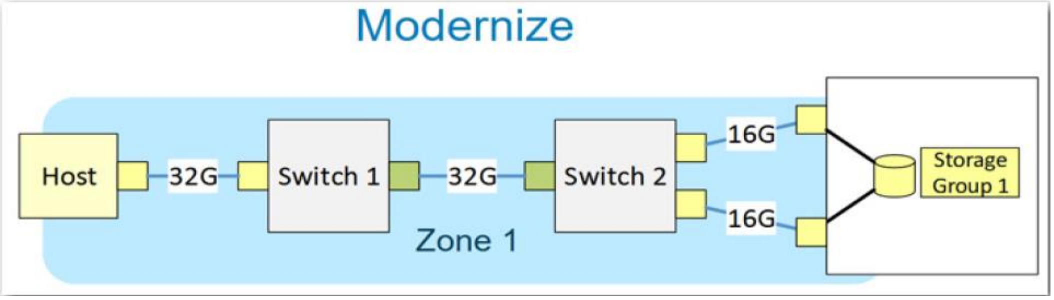


Figure 14 Modernize

Another way would be to re-zone as shown Figure 15.

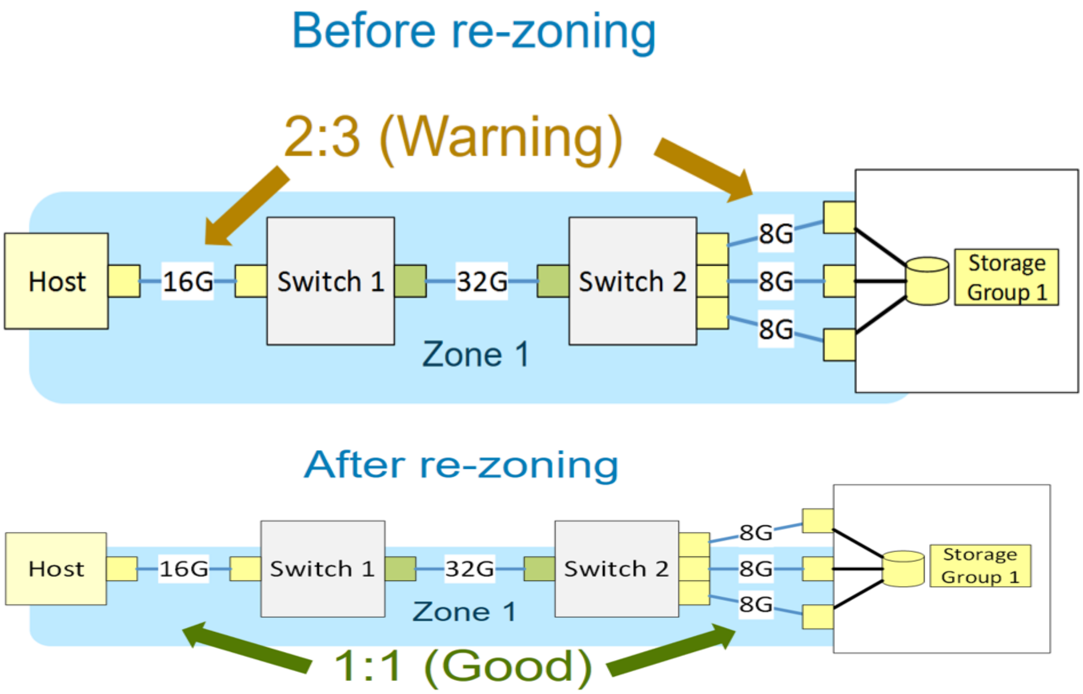


Figure 15 Before and After Re-zoning

Implement Bandwidth Limits

On Dell-EMC VMAX and Unity platforms create bandwidth limits on the Storage Groups (VMAX) or the LUNs (Unity). In the case study above where we had congestion spreading due to oversubscription when we implemented bandwidth limits we saw the performance was restored as noted in Figure 16 below. This can be done directly through Unisphere on the Storage group.

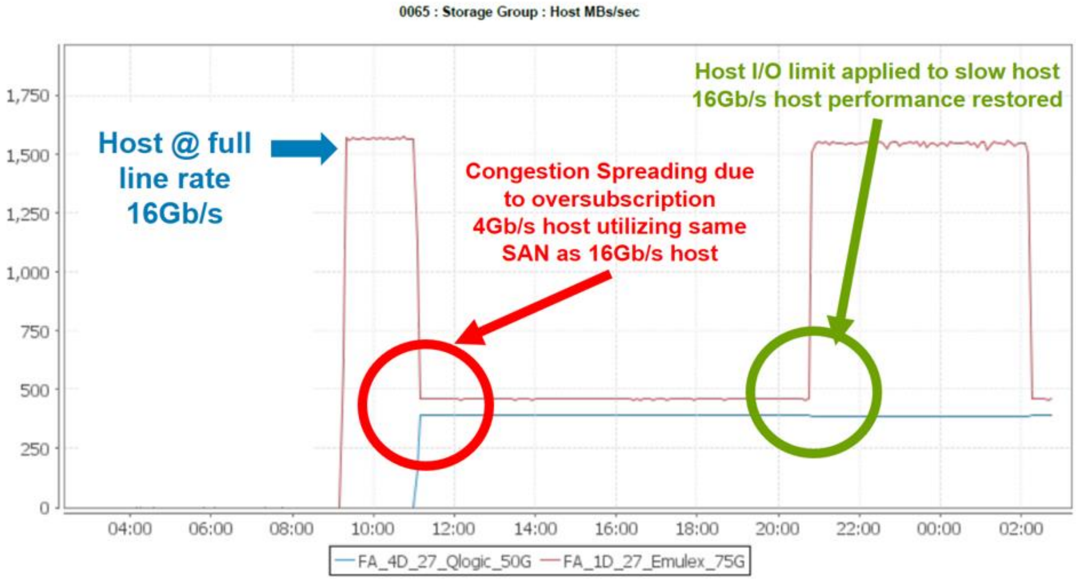


Figure 16 Host I/O Limited Applied

With I/O limits, it's important to note that this will not work well with clusters. Let's take Figure 17, below, for example. When the host limit is applied to a 4Gb host that is causing the back pressure, the array starts to limit the amount of data it sends back to the 4Gb (based on the IO limit set) therefore you eliminate the back-pressure issue all together and other flows can operate at full line rate.

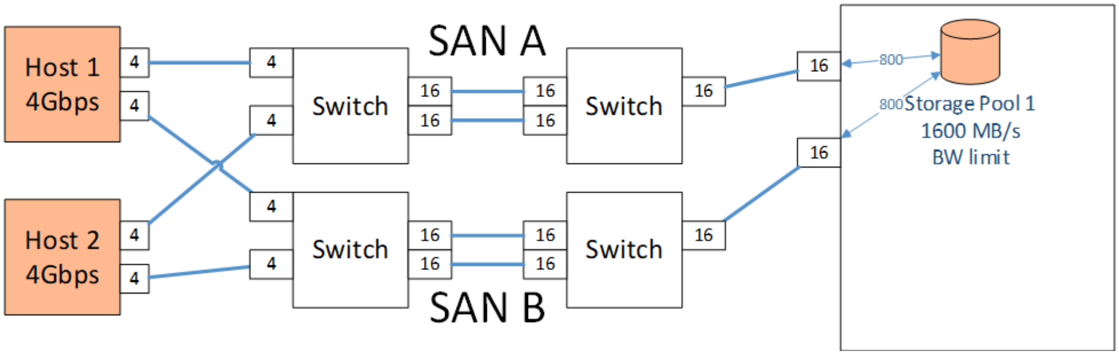


Figure 17 I/O Limits with clusters

In this example, we have two host running at 4Gb/s that are in a cluster. Because they are in a cluster both host will have access to the volume via each fabric, so that means we need to set a bandwidth I/O limit of 1600MB/s (800MB/s for each FA). However, with this approach there is nothing prevents a single HBA from consuming all 800MB/s.

- Isolation

Another way to prevent this issue is to isolate your slower traffic from your high-speed traffic and using dedicated ISLs. This can be achieved by creating Virtual Fabrics (Brocade) or VSANs (Cisco) as noted in Figure 18, below. The con with this approach is that you must dedicated ports, but this keeps your slower traffic from impacting your higher speed traffic. To enabled Virtual Fabrics on Brocade, it will require downtime, since the entire switch must be rebooted. When moving a port to a different VSAN on Cisco, only the end devices being moved are impacted.

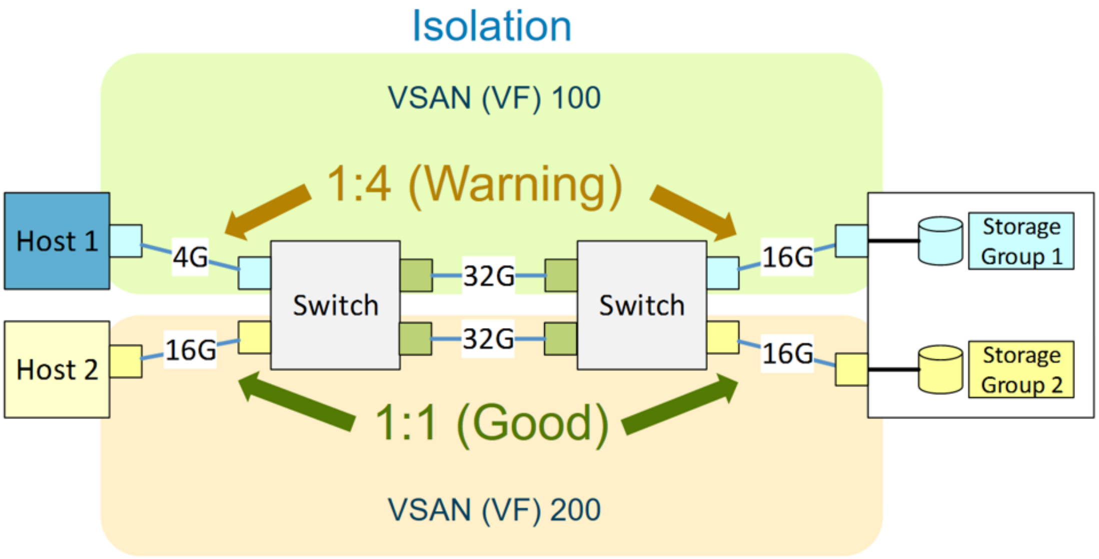


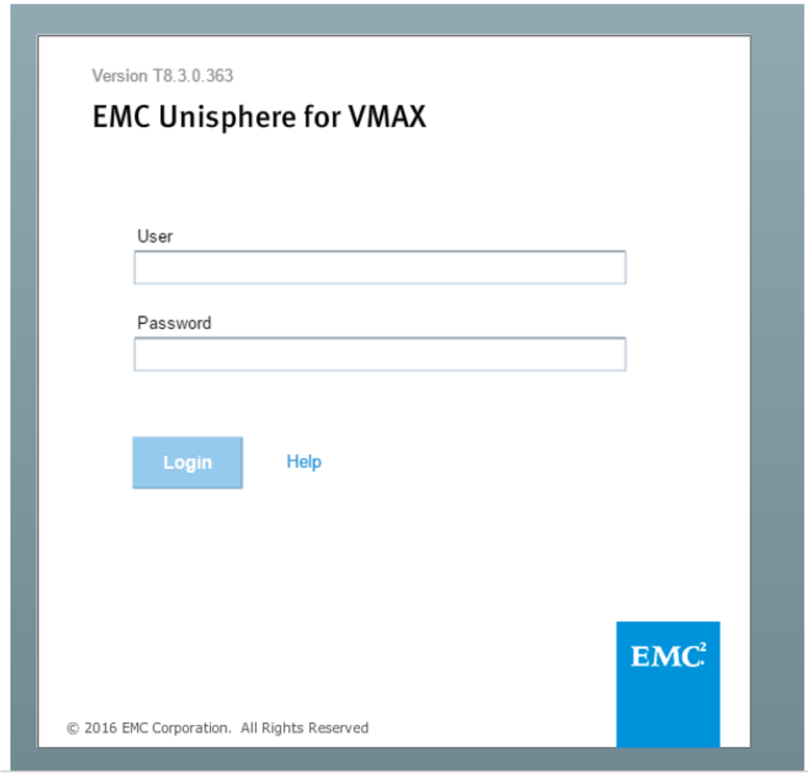
Figure 18 Isolation

6 Appendix

ENABLE PERFORMANCE MONITORING

This section provides steps on how to enable and review performance monitoring data in Unisphere for VMAX.

1. Login to the Unisphere GUI.



Version T8.3.0.363

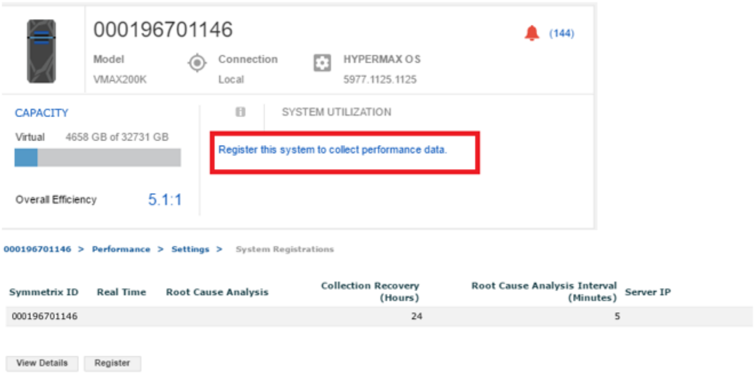
EMC Unisphere for VMAX

User

Password

© 2016 EMC Corporation. All Rights Reserved

2. Ensure the array is registered to collect performance data. If not, register the array.



000196701146

Model: VMAX200K

Connection: Local

HYPERMAX OS: 5977.1125.1125

CAPACITY

Virtual 4658 GB of 32731 GB

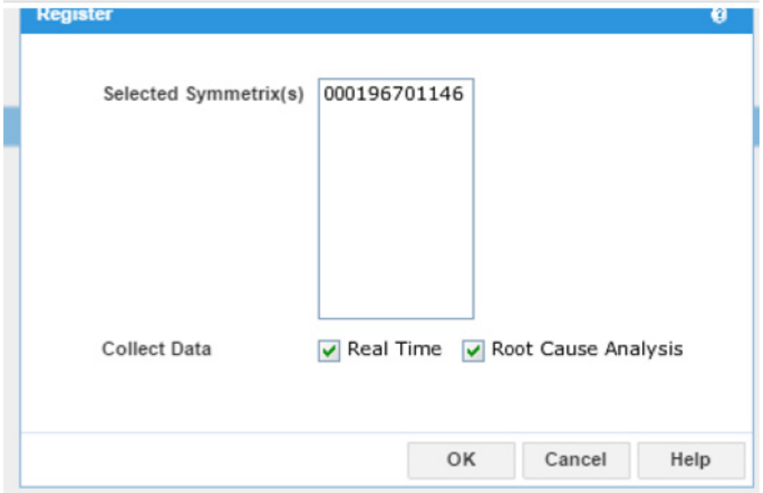
Overall Efficiency 5.1:1

SYSTEM UTILIZATION

Register this system to collect performance data.

000196701146 > Performance > Settings > System Registrations

Symmetrix ID	Real Time	Root Cause Analysis	Collection Recovery (Hours)	Root Cause Analysis Interval (Minutes)	Server IP
000196701146			24	5	



Register

Selected Symmetrix(s): 000196701146

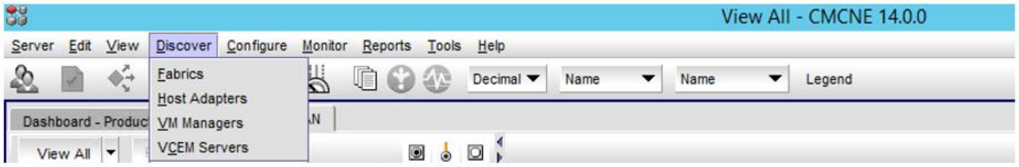
Collect Data: ☑ Real Time ☑ Root Cause Analysis

CONNECTRIX CONGESTION SPREADING MONITORING

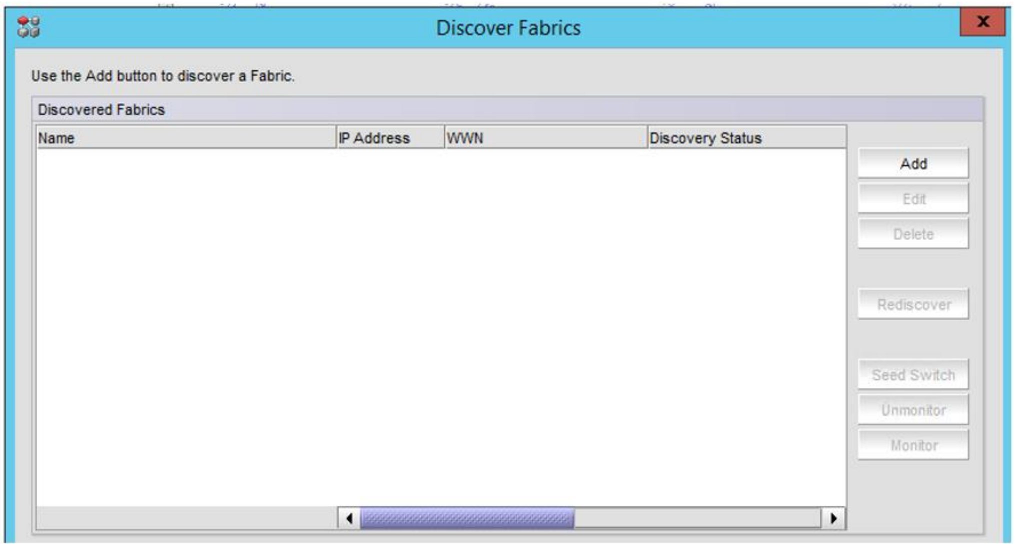
6.1.1 Brocade

- Discover the fabric

1. Login to the CMCNE Server and click **Discover >Fabrics**

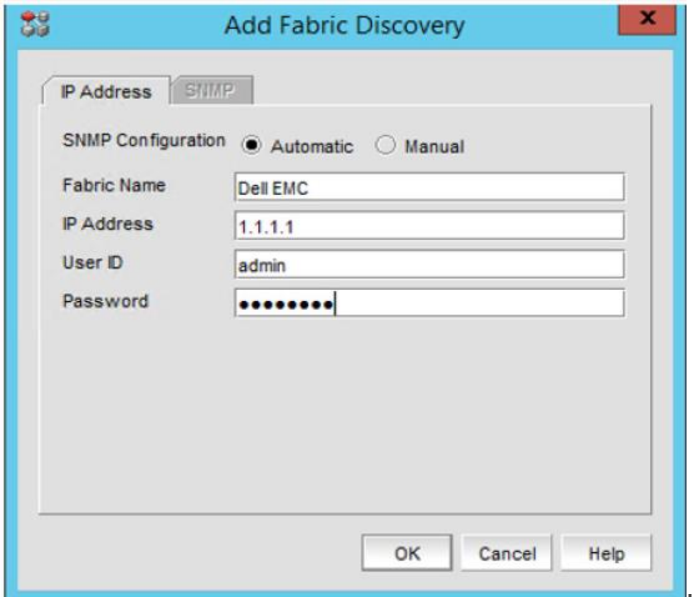


2. In the new window, click **Add.**



Use the Add button to discover a Fabric.

3. Fill in the required information for one of the switches in the fabric. CMCNE will automatically discover all switches in that fabric assuming the username and password are the same for all the switches in the fabric.



Add Fabric Discovery

SNMP Configuration ● Automatic ○ Manual

Fabric Name: Dell EMC

IP Address: 1.1.1.1

User ID: admin

Password: ••••••••

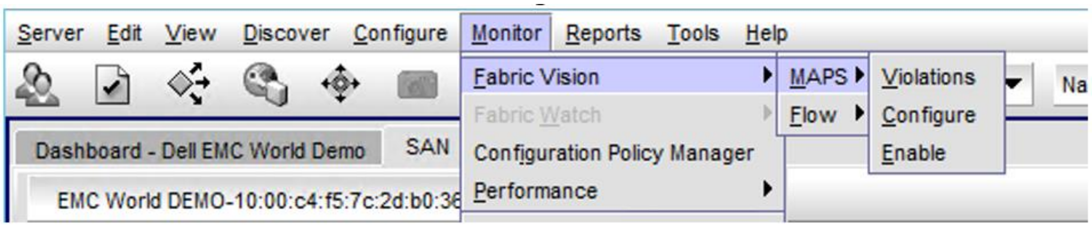
.

.

4. Repeat this section for all other fabrics.

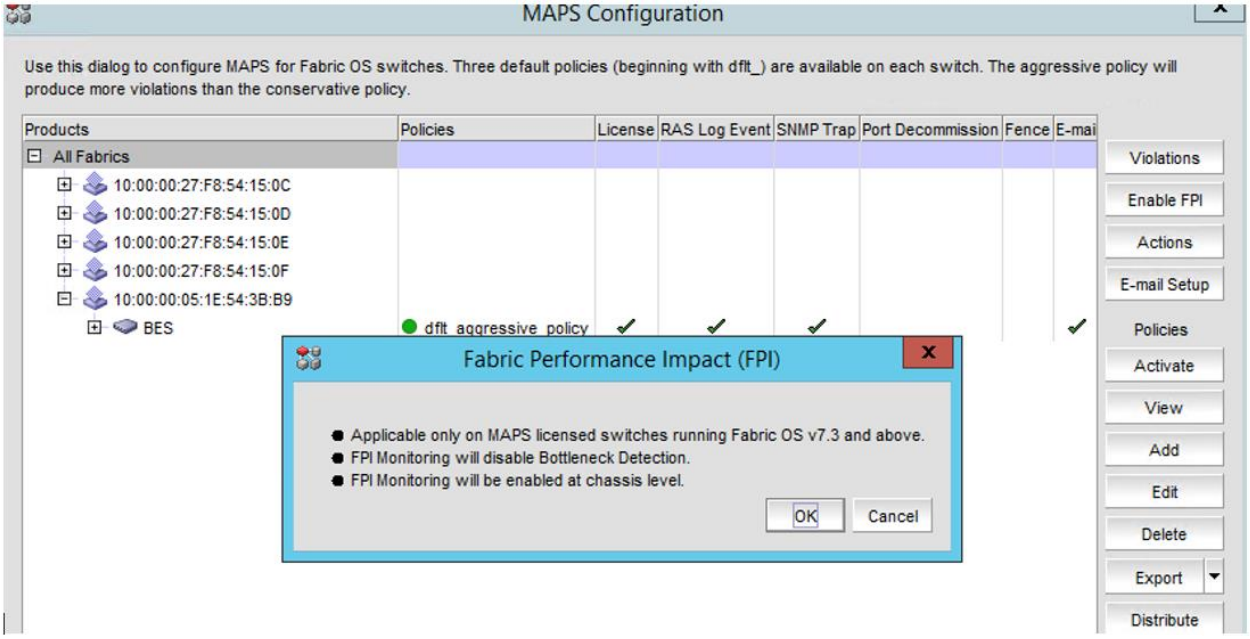
- Enable MAPS and FPI

1. Click on Monitor >Fabric Vision > MAPS > Configure



2. Highlight the fabric and enable FPI.

Note: FPI is enabled by default on switches running FOS 8.0 and above.



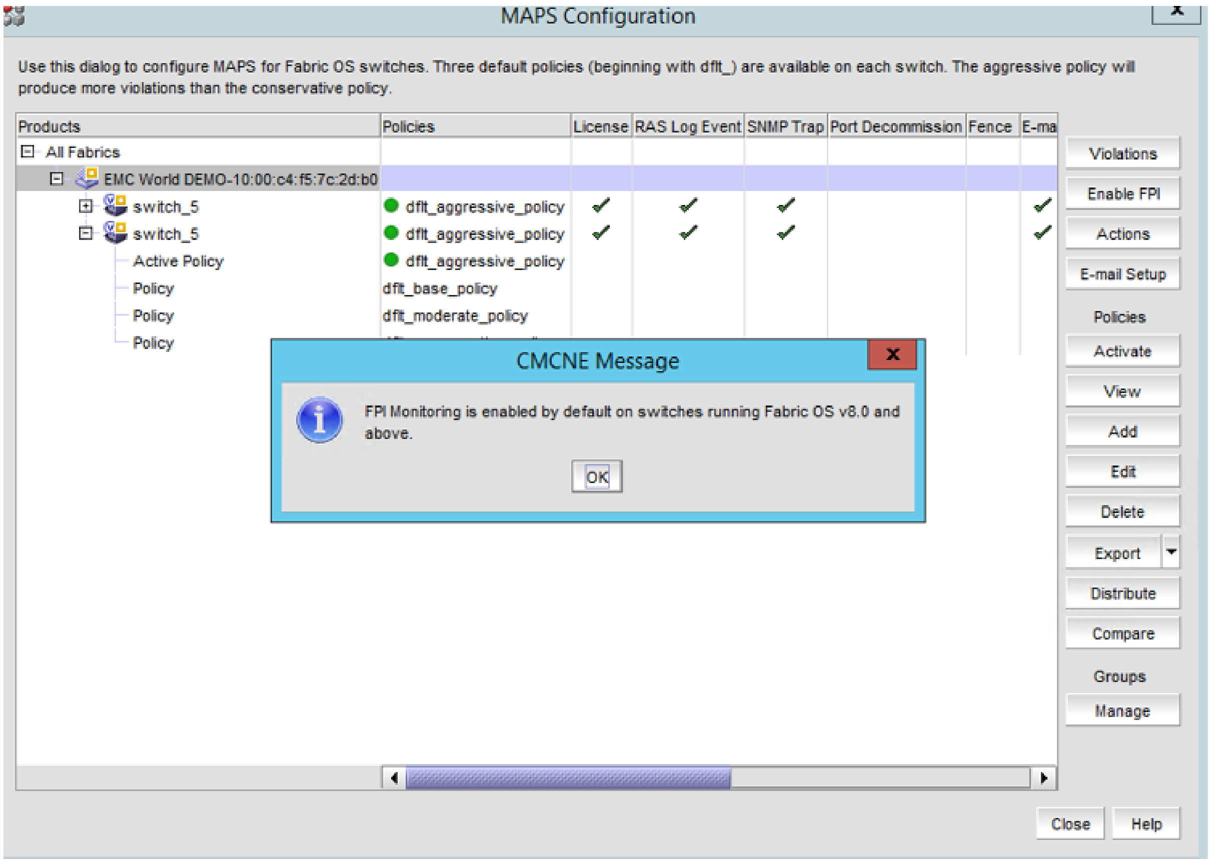
MAPS Configuration

Use this dialog to configure MAPS for Fabric OS switches. Three default policies (beginning with dflt_) are available on each switch. The aggressive policy will produce more violations than the conservative policy.

Products	Policies	License	RAS Log Event	SNMP Trap	Port Decommission	Fence	E-mai
⊟ All Fabrics							
⊞ 10:00:00:27:F8:54:15:0C							
⊞ 10:00:00:27:F8:54:15:0D							
⊞ 10:00:00:27:F8:54:15:0E							
⊞ 10:00:00:27:F8:54:15:0F							
⊟ 10:00:00:05:1E:54:3B:B9							
⊞ BES	● dflt_aggressive_policy	✓	✓	✓			✓

Fabric Performance Impact (FPI)

- Applicable only on MAPS licensed switches running Fabric OS v7.3 and above.
- FPI Monitoring will disable Bottleneck Detection.
- FPI Monitoring will be enabled at chassis level.



MAPS Configuration

Use this dialog to configure MAPS for Fabric OS switches. Three default policies (beginning with dflt_) are available on each switch. The aggressive policy will produce more violations than the conservative policy.

Products	Policies	License	RAS Log Event	SNMP Trap	Port Decommission	Fence	E-ma
⊟ All Fabrics							
⊟ EMC World DEMO-10:00:c4:f5:7c:2d:b0							
⊞ switch_5	● dflt_aggressive_policy	✓	✓	✓			✓
⊟ switch_5	● dflt_aggressive_policy	✓	✓	✓			✓
Active Policy	● dflt_aggressive_policy						
Policy	dflt_base_policy						
Policy	dflt_moderate_policy						
Policy							

CMCNE Message

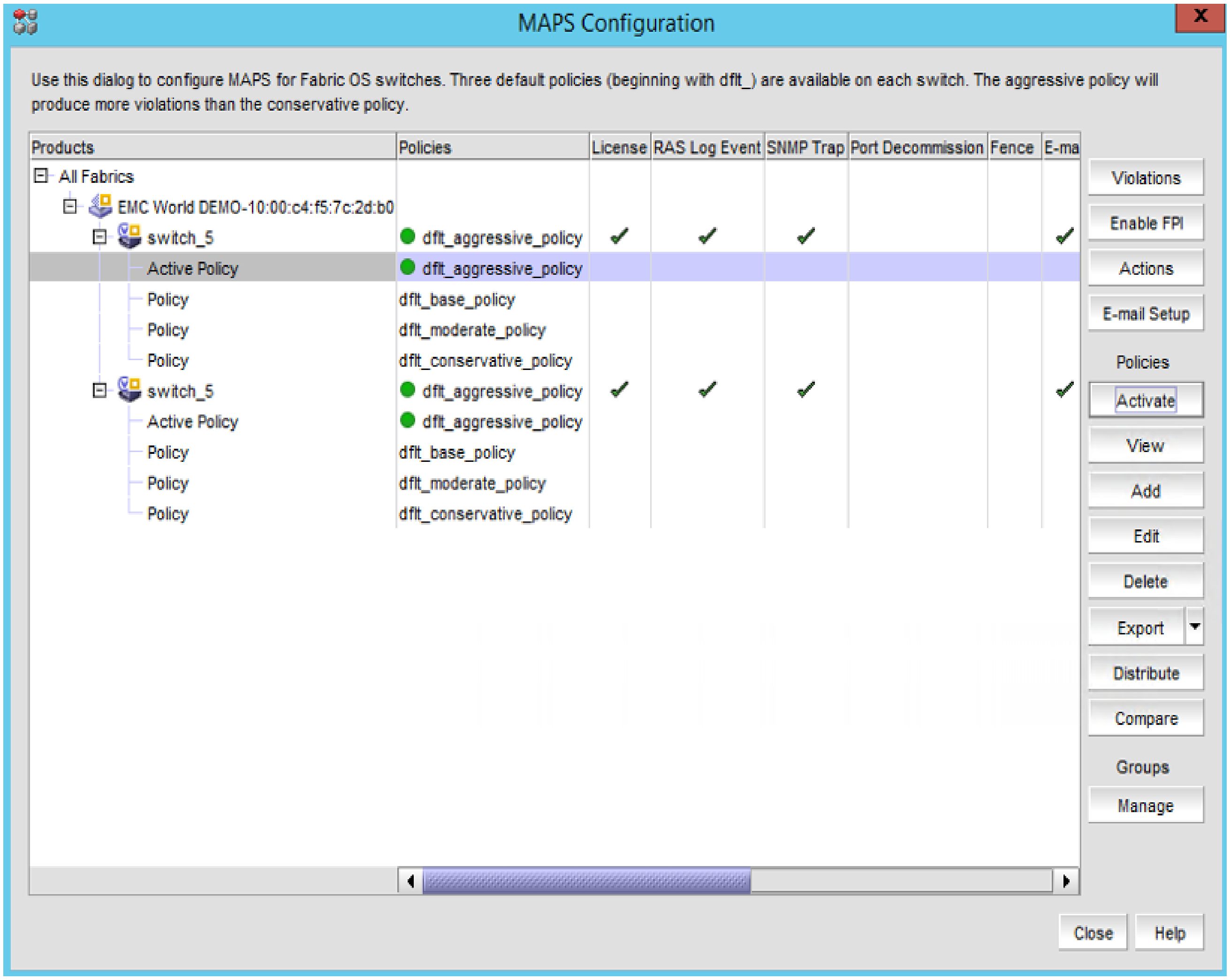
FPI Monitoring is enabled by default on switches running Fabric OS v8.0 and above.

3. From this menu you can configure each switch in your fabric and set the MAPS policy you want.

Note: CMCNE provides pre-defined policies that you can clone and then edit. You cannot edit the default policies. Reference the MAPS admin guide for further details on each policy and settings.

In this case we will activate the default aggressive policy. To do so, highlight **"dflt_aggressive_policy"** and click **activate**. This step must be repeated on ALL switches in the fabric that you want the policy enabled, currently you can't enable it for the entire fabric.

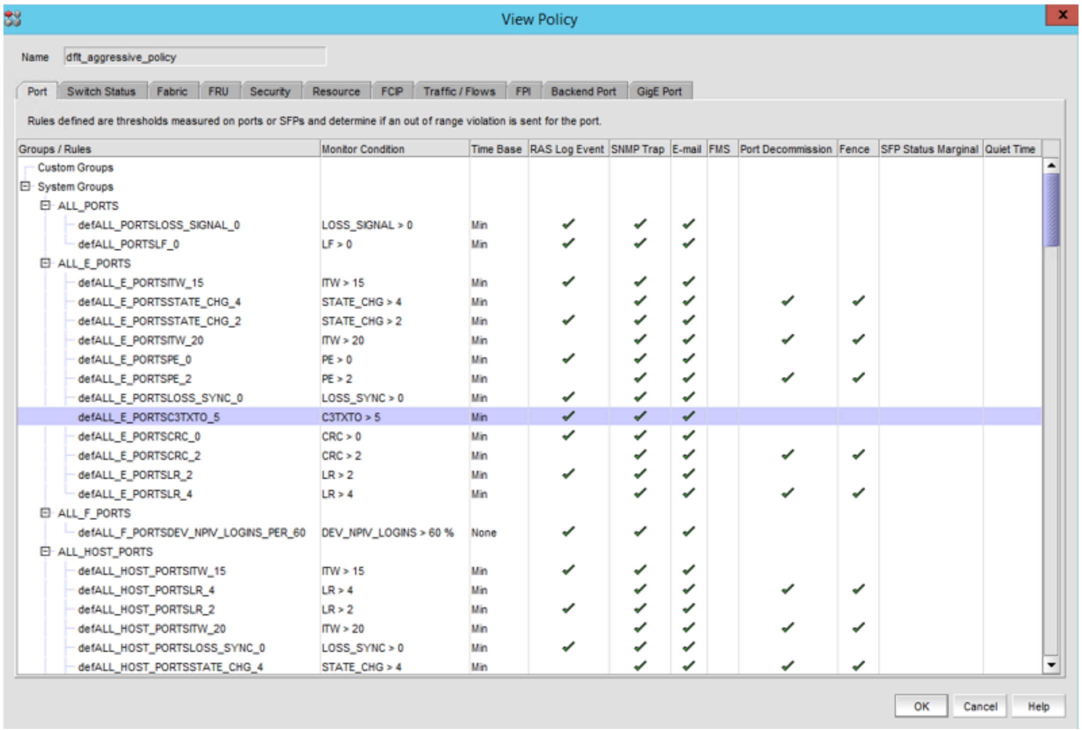
We are activating the aggressive policy first to get an idea of the issues in the fabric right away. After this you can adjust and use the other policies if we get too many alerts.
If you click on **View**, you can review the thresholds for each event.



MAPS Configuration

Use this dialog to configure MAPS for Fabric OS switches. Three default policies (beginning with dflt_) are available on each switch. The aggressive policy will produce more violations than the conservative policy.

Products	Policies	License	RAS Log Event	SNMP Trap	Port Decommission	Fence	E-ma
⊟ All Fabrics							
⊟ EMC World DEMO-10:00:c4:f5:7c:2d:b0							
⊟ switch_5	● dflt_aggressive_policy	✓	✓	✓			✓
Active Policy	● dflt_aggressive_policy						
Policy	dflt_base_policy						
Policy	dflt_moderate_policy						
Policy	dflt_conservative_policy						
⊟ switch_5	● dflt_aggressive_policy	✓	✓	✓			✓
Active Policy	● dflt_aggressive_policy						
Policy	dflt_base_policy						
Policy	dflt_moderate_policy						
Policy	dflt_conservative_policy						

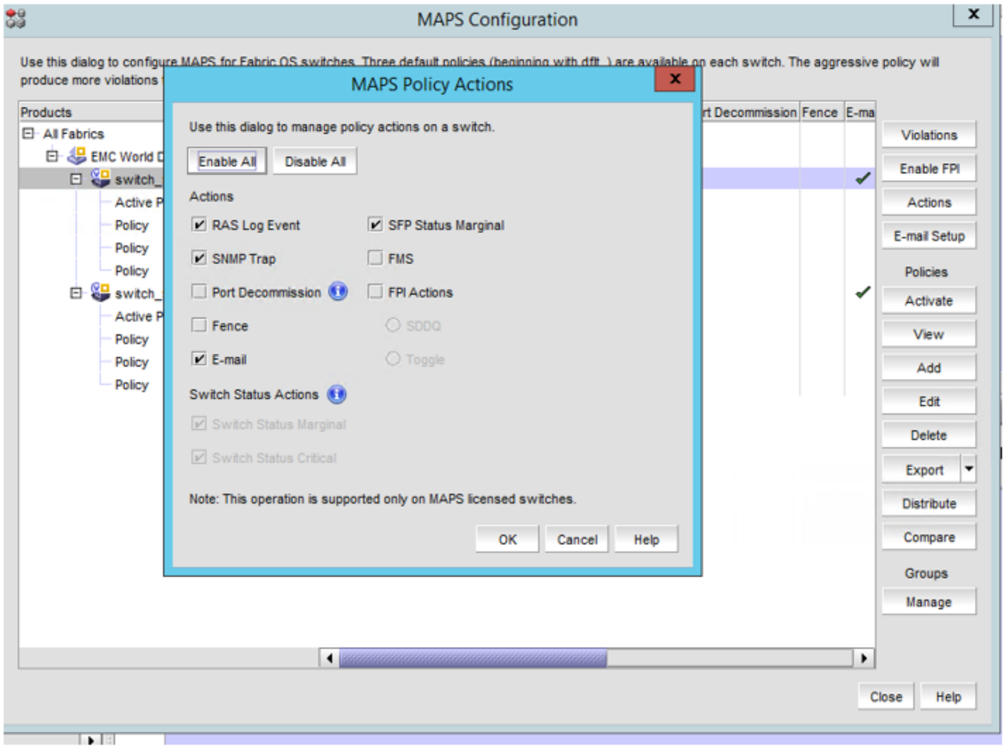


Name: dflt_aggressive_policy

Rules defined are thresholds measured on ports or SFPs and determine if an out of range violation is sent for the port.

Groups / Rules	Monitor Condition	Time Base	RAS Log Event	SNMP Trap	E-mail	FMS	Port Decommission	Fence	SFP Status Marginal	Quiet Time
Custom Groups										
System Groups										
ALL_PORTS										
defALL_PORTSLOSS_SIGNAL_0	LOSS_SIGNAL > 0	Min	✓	✓	✓					
defALL_PORTSLF_0	LF > 0	Min	✓	✓	✓					
ALL_E_PORTS										
defALL_E_PORTSITW_15	ITW > 15	Min	✓	✓	✓					
defALL_E_PORTSSTATE_CHG_4	STATE_CHG > 4	Min		✓	✓		✓	✓		
defALL_E_PORTSSTATE_CHG_2	STATE_CHG > 2	Min	✓	✓	✓					
defALL_E_PORTSITW_20	ITW > 20	Min		✓	✓		✓	✓		
defALL_E_PORTSPE_0	PE > 0	Min	✓	✓	✓					
defALL_E_PORTSPE_2	PE > 2	Min		✓	✓		✓	✓		
defALL_E_PORTSLOSS_SYNC_0	LOSS_SYNC > 0	Min	✓	✓	✓					
defALL_E_PORTSC3TXTO_5	C3TXTO > 5	Min	✓	✓	✓					
defALL_E_PORTSCRC_0	CRC > 0	Min	✓	✓	✓					
defALL_E_PORTSCRC_2	CRC > 2	Min		✓	✓		✓	✓		
defALL_E_PORTSLR_2	LR > 2	Min	✓	✓	✓					
defALL_E_PORTSLR_4	LR > 4	Min		✓	✓		✓	✓		
ALL_F_PORTS										
defALL_F_PORTSDEV_NPIV_LOGINS_PER_60	DEV_NPIV_LOGINS > 60 %	None	✓	✓	✓					
ALL_HOST_PORTS										
defALL_HOST_PORTSITW_15	ITW > 15	Min	✓	✓	✓					
defALL_HOST_PORTSLR_4	LR > 4	Min		✓	✓		✓	✓		
defALL_HOST_PORTSLR_2	LR > 2	Min	✓	✓	✓					
defALL_HOST_PORTSITW_20	ITW > 20	Min		✓	✓		✓	✓		
defALL_HOST_PORTSLOSS_SYNC_0	LOSS_SYNC > 0	Min	✓	✓	✓					
defALL_HOST_PORTSSTATE_CHG_4	STATE_CHG > 4	Min		✓	✓		✓	✓		

4. Highlight a switch and click on **Actions**. From here you can decide on the actions you would like to take in the event of congestion spreading. For our specific case study of congestion spreading due to oversubscription we would just need to ensure **email** and **RAS log event** are checked off.



Use this dialog to manage policy actions on a switch.

Actions

☑ RAS Log Event ☑ SFP Status Marginal

☑ SNMP Trap ☐ FMS

☐ Port Decommission ☐ FPI Actions

☐ Fence ○ SDDQ

☑ E-mail ○ Toggle

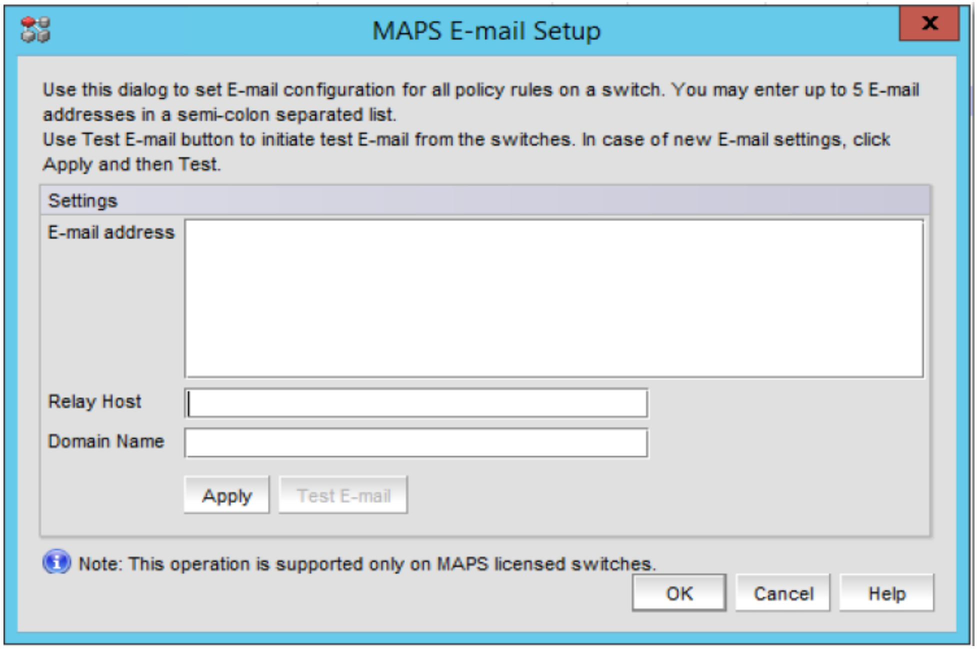
Switch Status Actions

☑ Switch Status Marginal

☑ Switch Status Critical

Note: This operation is supported only on MAPS licensed switches.

5. If you would like to receive alerts via email. Click on **E-mail Setup** and fill out the appropriate fields.



MAPS E-mail Setup

Use this dialog to set E-mail configuration for all policy rules on a switch. You may enter up to 5 E-mail addresses in a semi-colon separated list.
Use Test E-mail button to initiate test E-mail from the switches. In case of new E-mail settings, click Apply and then Test.

Settings

E-mail address

Relay Host

Domain Name

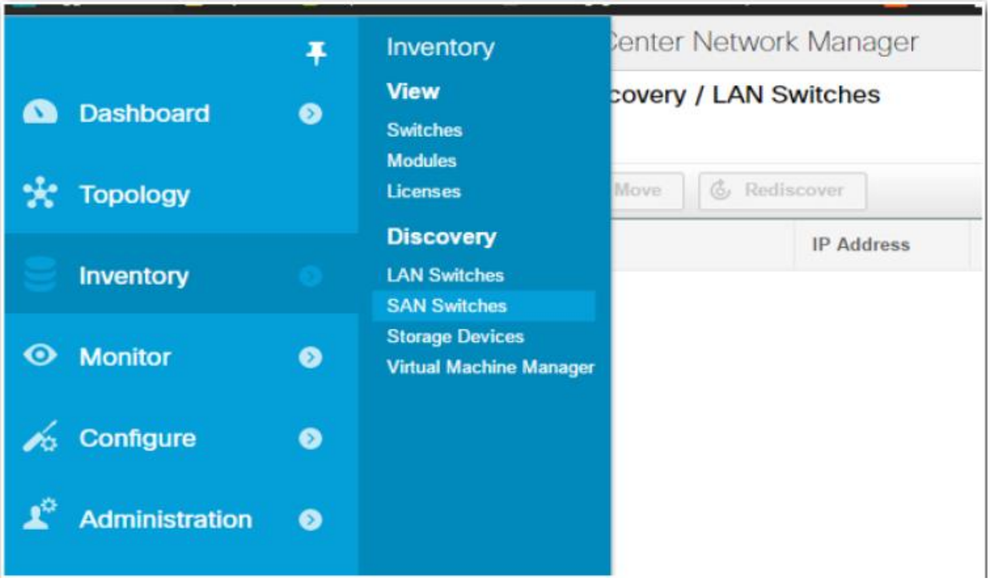
Note: This operation is supported only on MAPS licensed switches.

6. Ensure you repeated these steps on ALL Switches in the fabric.

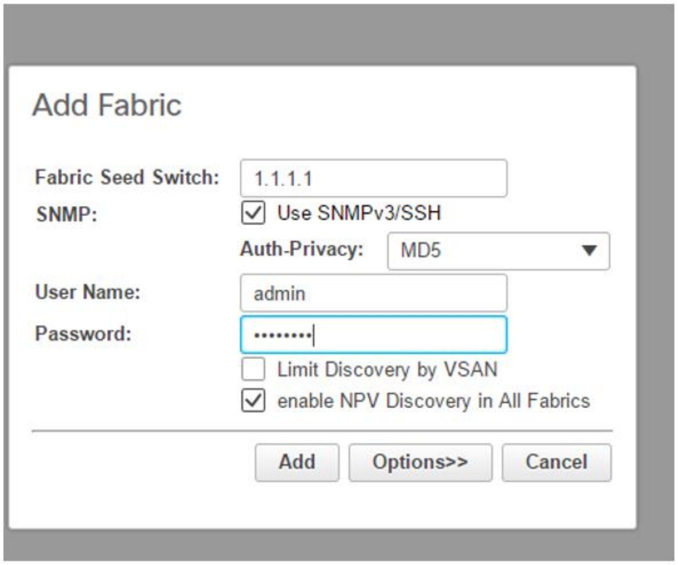
6.1.2 Cisco

- Discover the Fabric

1. Login to DCNM and click on Inventory>Discovery>SAN>Switches.



2. In the new window, click on **Plus Sign (+**) and fill in the required information for one of the switches in the fabric.



Add Fabric

Fabric Seed Switch: 1.1.1.1

SNMP: ☑ Use SNMPv3/SSH

Auth-Privacy: MD5

User Name: admin

Password: ••••••••

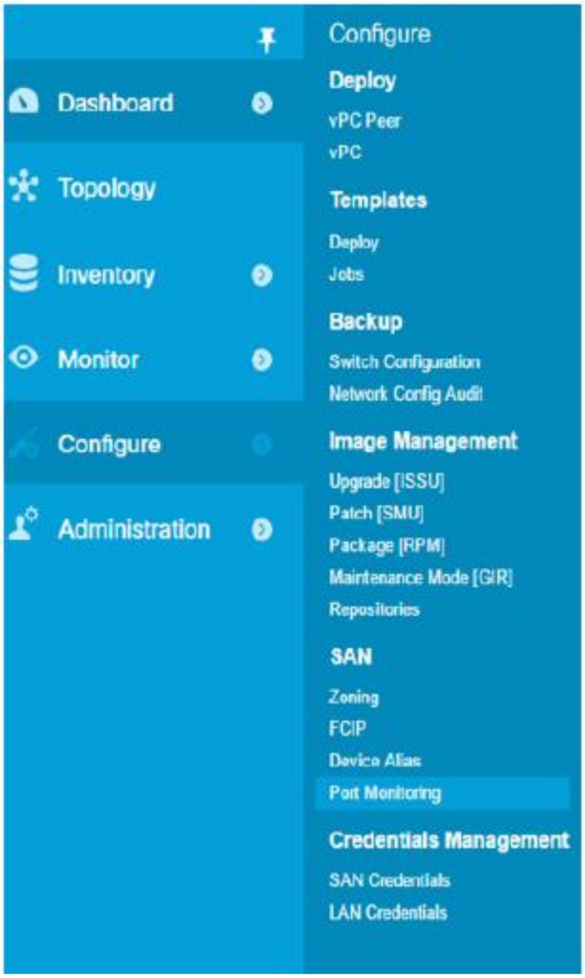
□ Limit Discovery by VSAN

☑ enable NPV Discovery in All Fabrics

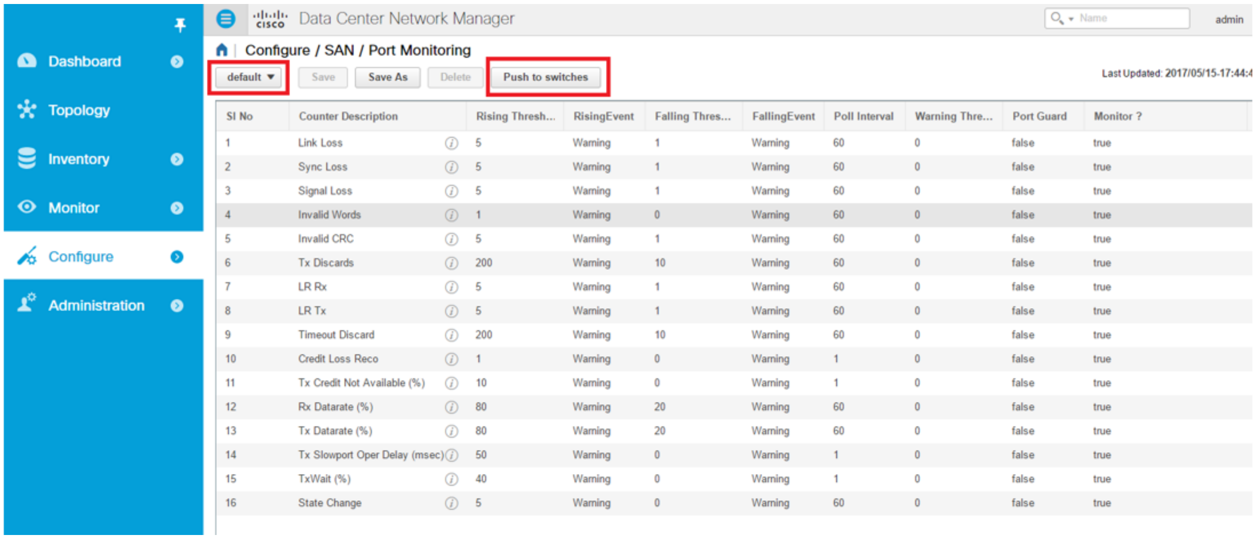
3. Repeat this section for all other fabrics.

- Enable Cisco Port Monitoring (PMON)

1. Click on **Configure > SAN > Port Monitoring**.



2. Select default profile and click **on Push to switches**.

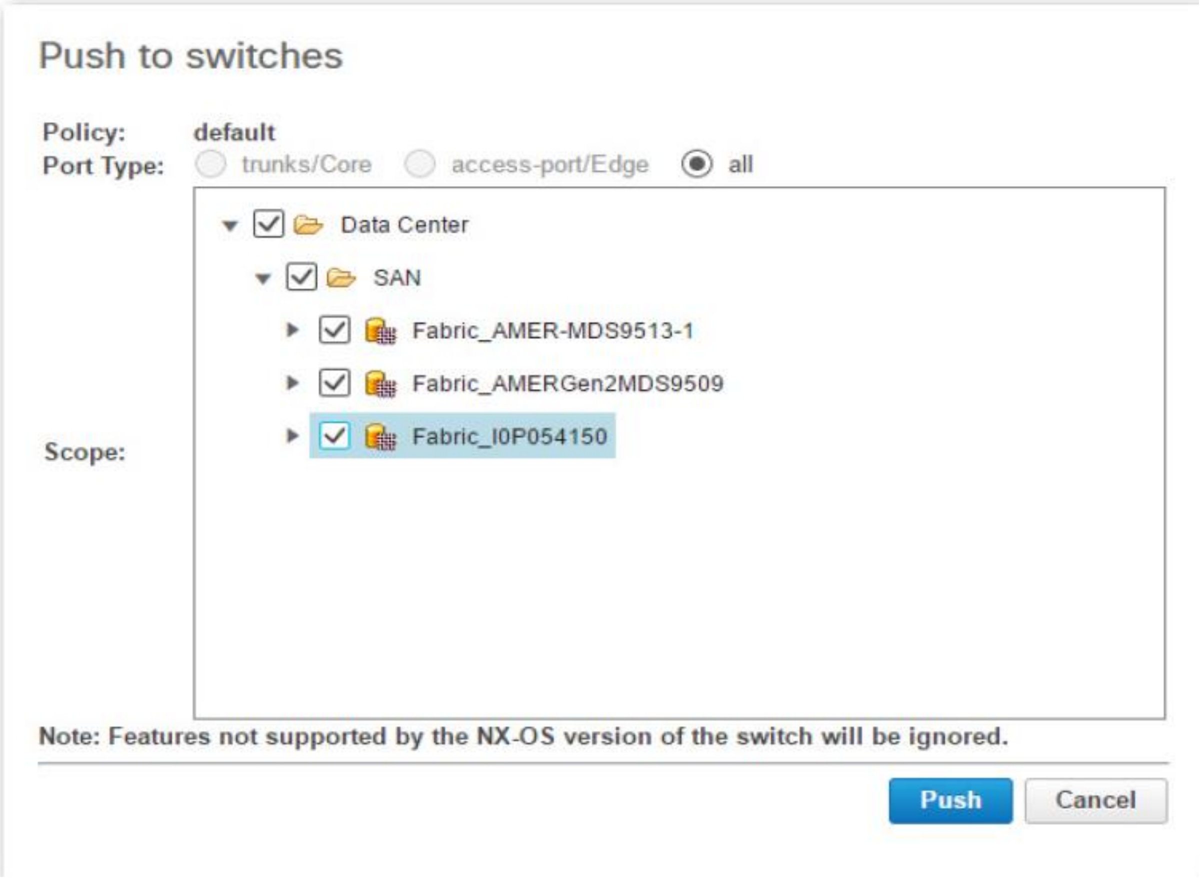


Configure / SAN / Port Monitoring

Last Updated: 2017/05/15-17:44:4

SI No	Counter Description	Rising Thresh...	RisingEvent	Falling Thres...	FallingEvent	Poll Interval	Warning Thre...	Port Guard	Monitor ?
1	Link Loss	5	Warning	1	Warning	60	0	false	true
2	Sync Loss	5	Warning	1	Warning	60	0	false	true
3	Signal Loss	5	Warning	1	Warning	60	0	false	true
4	Invalid Words	1	Warning	0	Warning	60	0	false	true
5	Invalid CRC	5	Warning	1	Warning	60	0	false	true
6	Tx Discards	200	Warning	10	Warning	60	0	false	true
7	LR Rx	5	Warning	1	Warning	60	0	false	true
8	LR Tx	5	Warning	1	Warning	60	0	false	true
9	Timeout Discard	200	Warning	10	Warning	60	0	false	true
10	Credit Loss Reco	1	Warning	0	Warning	1	0	false	true
11	Tx Credit Not Available (%)	10	Warning	0	Warning	1	0	false	true
12	Rx Datarate (%)	80	Warning	20	Warning	60	0	false	true
13	Tx Datarate (%)	80	Warning	20	Warning	60	0	false	true
14	Tx Slowport Oper Delay (msec)	50	Warning	0	Warning	1	0	false	true
15	TxWait (%)	40	Warning	0	Warning	1	0	false	true
16	State Change	5	Warning	0	Warning	60	0	false	true

3. Select all the fabrics and click **Push**.



Push to switches

Policy: default

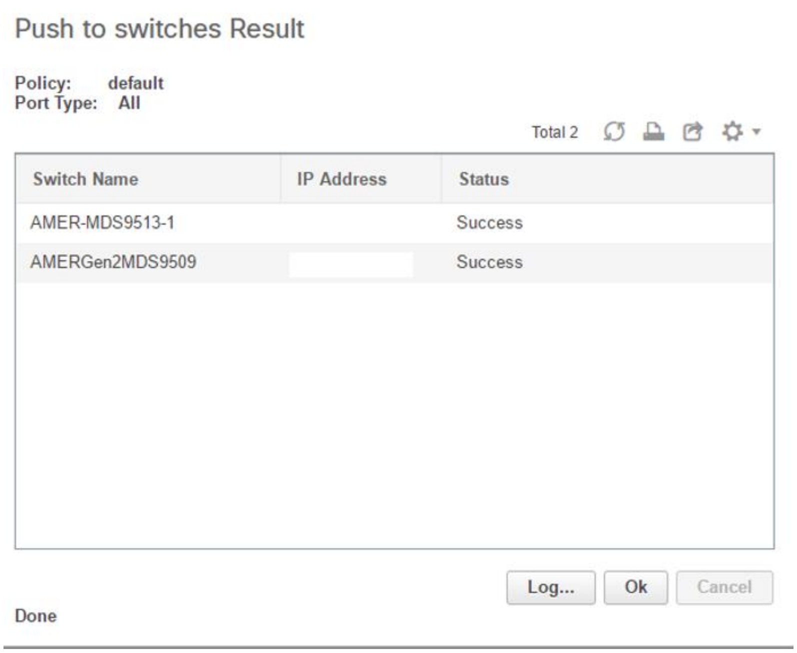
Port Type: ○ trunks/Core ○ access-port/Edge ● all

Scope:

- ☑ Data Center
  - ☑ SAN
    - ☑ Fabric_AMER-MDS9513-1
    - ☑ Fabric_AMERGen2MDS9509
    - ☑ Fabric_I0P054150

Note: Features not supported by the NX-OS version of the switch will be ignored.

Note: The IP addresses were removed on purpose.



Push to switches Result

Policy: default
Port Type: All

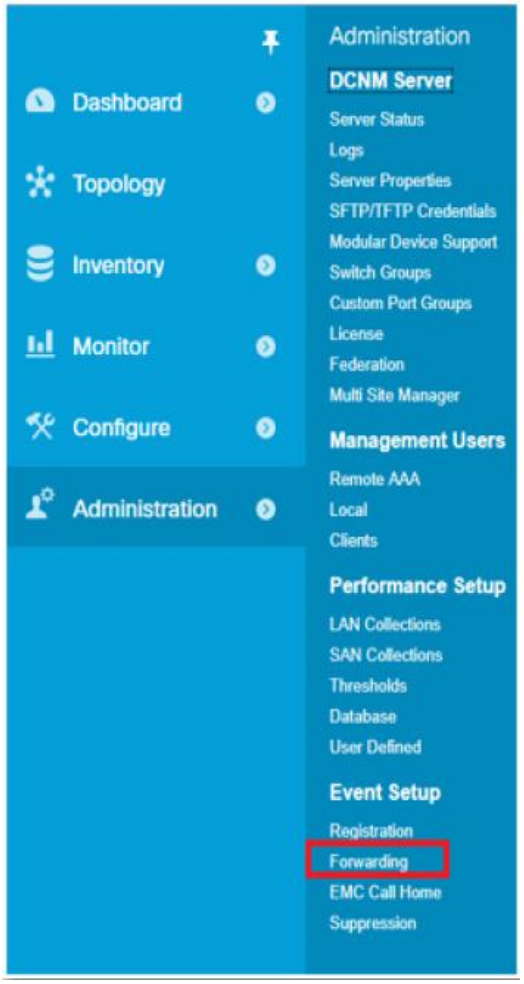
Switch Name	IP Address	Status
AMER-MDS9513-1		Success
AMERGen2MDS9509		Success

Done

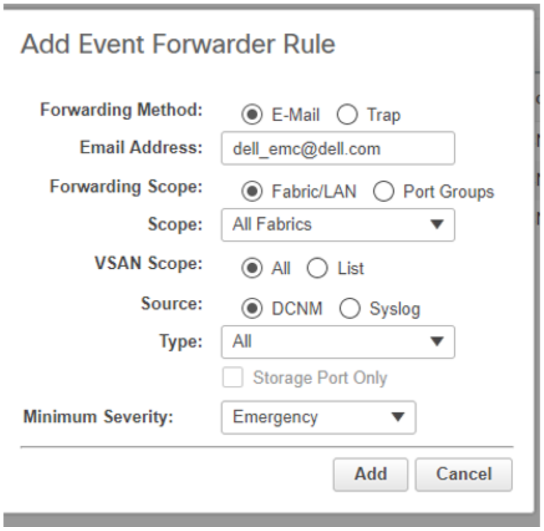
4. With Cisco MDS, you can receive the alerts via SNMP or Syslog. Please refer to the following configuration guide for configuration on both:

   http://www.cisco.com/c/en/us/support/storage-networking/mds-9000-nx-os-san-os-software/products-installation-and-configuration-guides-list.html

5. To configure Email home (optional), click on **Administration > Event Setup.**



6. Click on the **Plus (+) sign,** provide the recipient email address, and click **Add.**



Add Event Forwarder Rule

Forwarding Method: ● E-Mail ○ Trap

Email Address: dell_emc@dell.com

Forwarding Scope: ● Fabric/LAN ○ Port Groups

Scope: All Fabrics

VSAN Scope: ● All ○ List

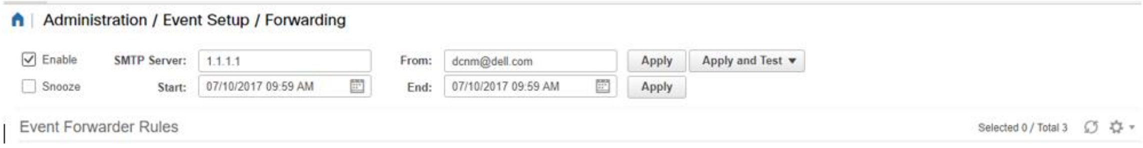
Source: ● DCNM ○ Syslog

Type: All

□ Storage Port Only

Minimum Severity: Emergency

7. Fill out the SMTP Server information and the sender email address, then hit **apply and test** to confirm that you received the email.



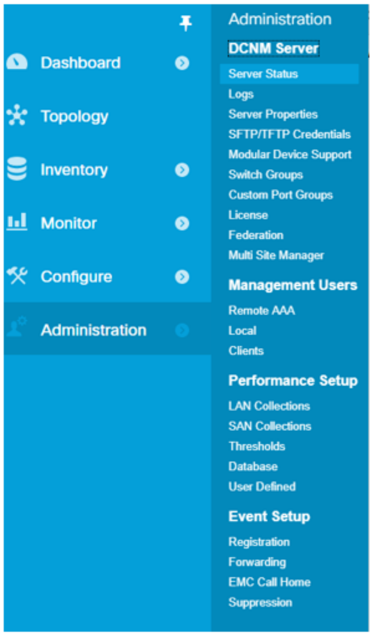
Administration / Event Setup / Forwarding

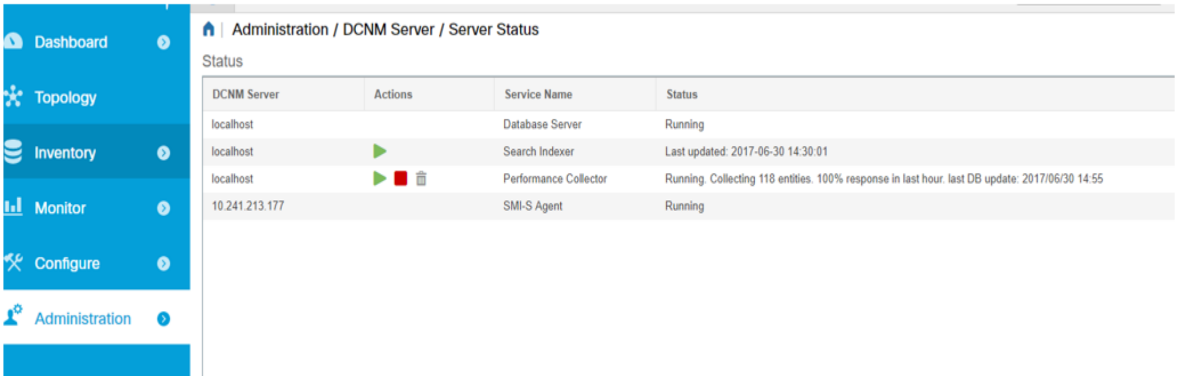
☑ Enable SMTP Server: 1.1.1.1 From: dcnm@dell.com Apply Apply and Test ▾

□ Snooze Start: 07/10/2017 09:59 AM End: 07/10/2017 09:59 AM Apply

Event Forwarder Rules Selected 0 / Total 3

8. Ensure Performance Monitoring is running. Click on **Administration >Server Status**. Ensure that the **Performance Collector** is running. If not, hit the **play** button to start it.



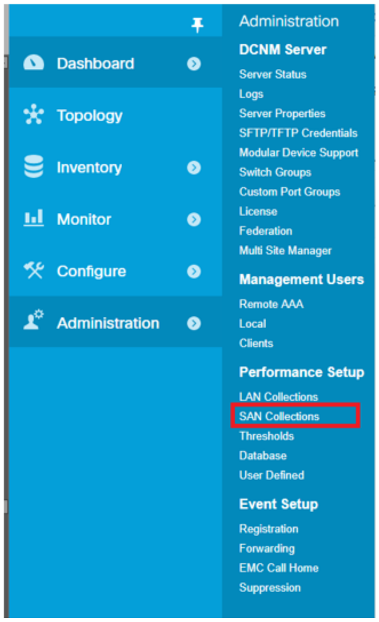


Administration / DCNM Server / Server Status

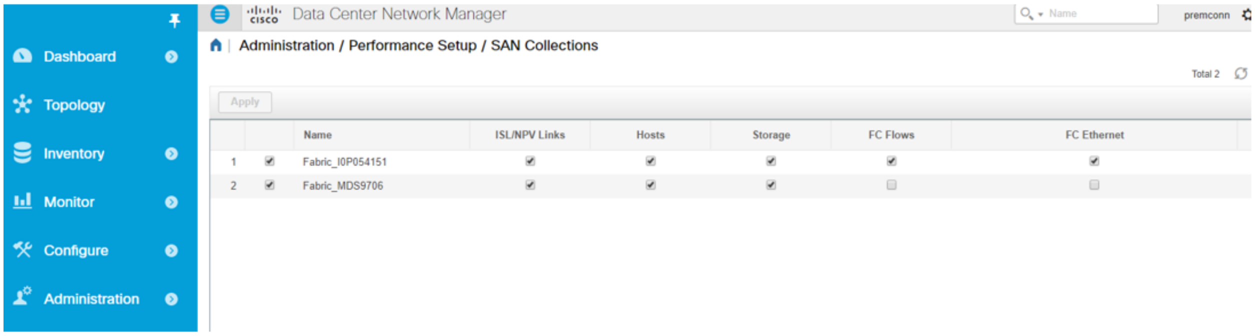
Status

DCNM Server	Actions	Service Name	Status
localhost		Database Server	Running
localhost	▶	Search Indexer	Last updated: 2017-06-30 14:30:01
localhost	▶ ■ 🗑	Performance Collector	Running. Collecting 118 entities. 100% response in last hour. last DB update: 2017/06/30 14:55
10.241.213.177		SMI-S Agent	Running

9. Click on **Administration >Performance Setup >SAN> Collections.**



10. Ensure the fabric(s) for which you want to collect the performance statistics are checked. The Performance Collector service will restart.



Administration / Performance Setup / SAN Collections

Total 2

Apply

		Name	ISL/NPV Links	Hosts	Storage	FC Flows	FC Ethernet
1	☑	Fabric_I0P054151	☑	☑	☑	☑	☑
2	☑	Fabric_MDS9706	☑	☑	☑	□	□

References

1. Brocade MAPS Configuration Guide:

   http://www.brocade.com/content/html/en/configuration-guide/fos-80x-maps/GUID-426E1CD4-3763-419D-9D54-91F824F463EB-homepage.html

2. Cisco Slow Drain Device White Paper:

   http://www.cisco.com/c/dam/en/us/products/collateral/storage-networking/mds-9700-series-multilayer-directors/whitepaper-c11-737315.pdf

General reference about VMAX Host I/0 Limits features:

https://community.emc.com/thread/188068?start=0&tstart=0

4. Ezfio I/O tool

https://github.com/earlephilhower/ezfio

Congestion Spreading Severity

While the congestion spreading metrics are all important, as the following section will illustrate, the rate at which the events are occurring can dramatically alter the impact that each event can have on your environment. Further complicating matters is the fact that both Brocade and Cisco use a different severity categorization scheme. As a result, we will use the following Dell EMC specific categorization scheme and map them to each of the switch types as shown below:

6.1.3 Dell EMC

- Type 1:
  - Congestion Ratio greater than or equal to .2
  - No frame loss (discards) or link resets
- Type 2:
  - Congestion Ratio greater than or equal to .2
  - Frame loss (discards), but no link resets
- Type 3:
  - Congestion Ratio greater than or equal to .2
  - Frame loss (discards) and link resets

6.1.4 Brocade

- Mild
  - Small credit delay
  - Small queue latency (less than 10ms)
  - No frame loss (discards) or link resets
- Moderate
  - Medium credit delay
  - Medium queue latency (10ms – 80ms)
  - Frame loss (discards), but no link resets
- Severe
  - Large credit delay

  - Large queue latency (greater than 80ms)
  - Frame loss (discards) and some link resets

6.1.5 Cisco

- Level – 1: Latency

  - Reduced number of remaining credits or small duration of credit unavailability
  - No discards, retransmission or link resets

- Level – 2: Retransmission

  - Longer duration of credit unavailability
  - Frames are discarded (but no link reset) due to congestion-drop timeout or no-credit-drop timeout* leading to retransmission.

- Level – 3: Extreme delay

  - Prolonged duration of credit unavailability (1 sec for F-port, 1.5 sec for E-port)
  - Link resets or port flaps

Congestion Spreading Terminology Cross Reference

The metrics and severities can be combined and used to help identify the different types of congestion spreading events. As with the previous section, there's a separate section for both Brocade and
Cisco. However, since both Brocade and Cisco use the term oversubscription, this section will start with an overview of it first.

6.1.6 Oversubscription

In simplest terms, oversubscription is a condition where "the potential demand on a system exceeds the capacity of the system to meet that demand." A very simple example that most are familiar with is the highway system. If everyone suddenly decided to drive their cars at the same time (such as during a hurricane evacuation event), the traffic would be at a standstill.

In the case of a FC SAN, it's useful to think of oversubscription in terms of a bandwidth (BW) ratio. For example, as shown in figure 3, the BW ratio between Host 1 (4Gbps) and Storage 1 (16 Gbps) is 1:4.
We can therefore say that Host 1 is 4:1 oversubscribed. Compare this with the BW ratio between Host 2 (16 Gbps) and Storage 2 (16Gbps) which is 1:1, and consider that both hosts and the storage they access
will utilize a 32 Gbps ISL, and you can see that there is no oversubscription between Host 2 and Storage 2. In this case we say that Host 2 and Storage 2 are non-oversubscribed.

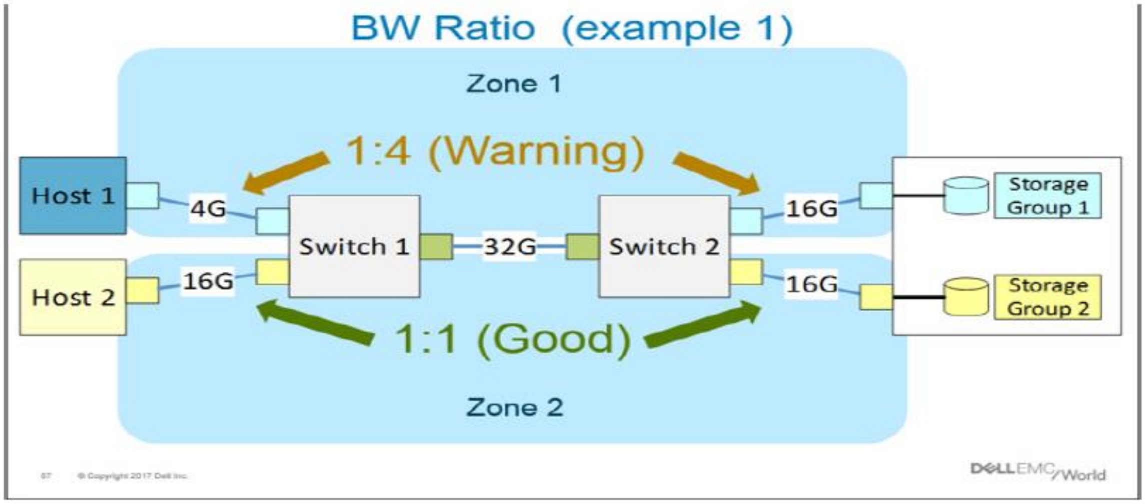
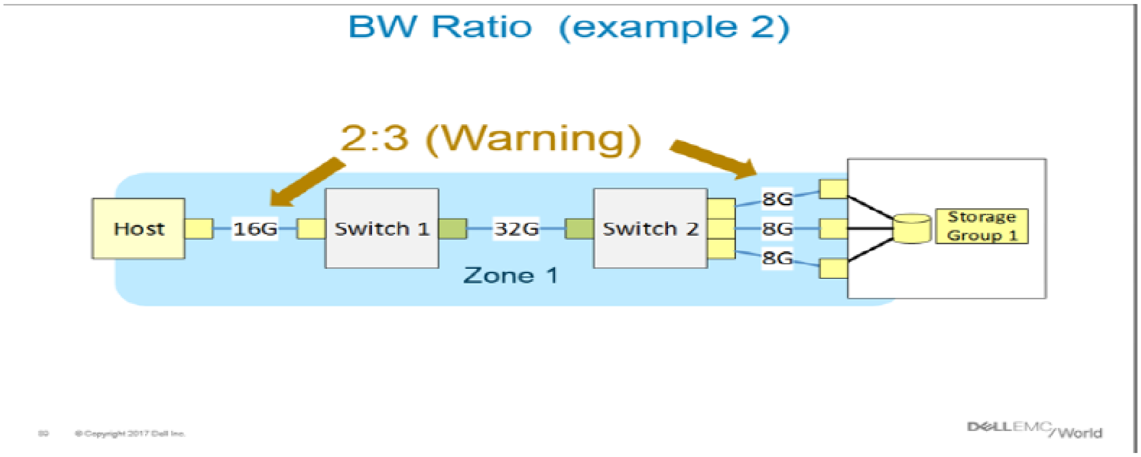


Figure 19 Bandwidth Ratio - Example 1

It's important to note that when calculating oversubscription, as shown in Figure 19 the bandwidth ratio is calculated by summing the BW of the interfaces that are being considered. At first glance, you
might think that we have a 16 Gbps HBA accessing 8 Gbps storage, but since in fact there are three storage interfaces, we have a 16 Gbps HBA accessing 24 Gbps of storage and as a result the host is 3:2 oversubscribed.

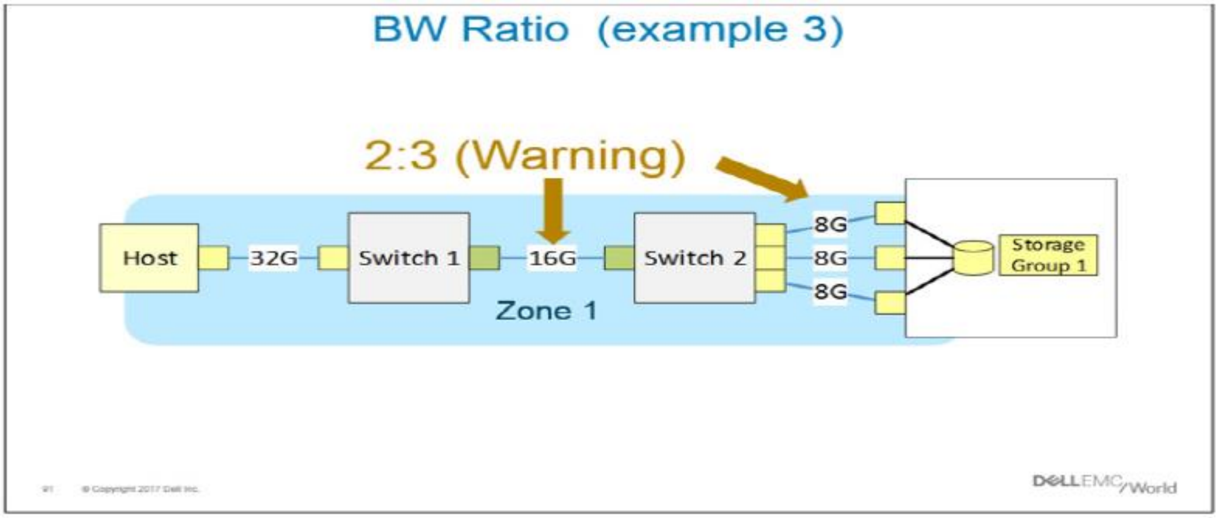


BW Ratio (example 2)

2:3 (Warning)

Figure 20 BW Ratio - Example 2

In the previous 2 examples, the BW of the ISL was always greater than or equal to the amount of BW that the end devices could support. This will not normally be the case. As shown in Figure 20, the host is in fact 3:4 undersubscribed, but since the ISL is only 16 Gbps, there is oversubscription between the end device and the ISLs that will be used and you could say the ISLs are 3:2 oversubscribed.



BW Ratio (example 3)

2:3 (Warning)

Figure 21 BW Ratio - Example 3

6.1.7 Brocade

Brocade defines three different classes of congestion spreading events:

- Oversubscription

As defined in the preceding section (above).

- Misbehaving Device

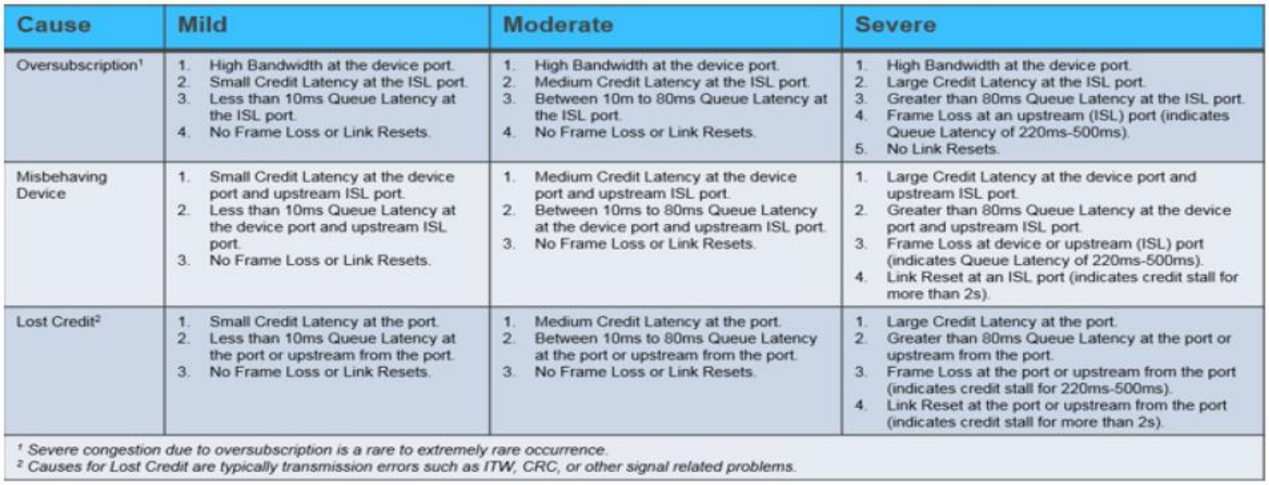
An end device or ISL that is not releasing credit at a rate fast enough to sustain line rate. For example, if an end device has negotiated a link speed of 16 Gbps and is unable to return credit at a rate that allows it to receive 16Gbps of data, then you could have a misbehaving device. These types of devices are also referred to as "slow drains". It's important to point out that a device might be misbehaving for a number of reasons including a driver problem or in the case of a ISL, because the port is experiencing the effects of congestion spreading.

- Lost Credit

A Lost Credit scenario means that for one reason or another (e.g.: usually occasional bit errors), one or both devices on a given link believe that they have fewer transmit credit than they really do. One cause of this situation would be for a bit error to corrupt an R_RDY. If this happens often enough, the performance will start to drop off over time and slowly degrade the ability of the SAN to transport data. This issue is explored in more detail in KB 464245 (Bit Errors and their impact).

Brocade Congestion Spreading Terminology Cross Reference

For Brocade, bringing all the pieces together results in the following Brocade-specific Congestion Spreading Terminology cross reference.



Cause	Mild	Moderate	Severe
Oversubscription[1]	1. High Bandwidth at the device port. 2. Small Credit Latency at the ISL port. 3. Less than 10ms Queue Latency at the ISL port. 4. No Frame Loss or Link Resets.	1. High Bandwidth at the device port. 2. Medium Credit Latency at the ISL port. 3. Between 10m to 80ms Queue Latency at the ISL port. 4. No Frame Loss or Link Resets.	1. High Bandwidth at the device port. 2. Large Credit Latency at the ISL port. 3. Greater than 80ms Queue Latency at the ISL port. 4. Frame Loss at an upstream (ISL) port (indicates Queue Latency of 220ms-500ms). 5. No Link Resets.
Misbehaving Device	1. Small Credit Latency at the device port and upstream ISL port. 2. Less than 10ms Queue Latency at the device port and upstream ISL port. 3. No Frame Loss or Link Resets.	1. Medium Credit Latency at the device port and upstream ISL port. 2. Between 10ms to 80ms Queue Latency at the device port and upstream ISL port. 3. No Frame Loss or Link Resets.	1. Large Credit Latency at the device port and upstream ISL port. 2. Greater than 80ms Queue Latency at the device port and upstream ISL port. 3. Frame Loss at device or upstream (ISL) port (indicates Queue Latency of 220ms-500ms). 4. Link Reset at an ISL port (indicates credit stall for more than 2s).
Lost Credit[2]	1. Small Credit Latency at the port. 2. Less than 10ms Queue Latency at the port or upstream from the port. 3. No Frame Loss or Link Resets.	1. Medium Credit Latency at the port. 2. Between 10ms to 80ms Queue Latency at the port or upstream from the port. 3. No Frame Loss or Link Resets.	1. Large Credit Latency at the port. 2. Greater than 80ms Queue Latency at the port or upstream from the port. 3. Frame Loss at the port or upstream from the port (indicates credit stall for 220ms-500ms). 4. Link Reset at the port or upstream from the port (indicates credit stall for more than 2s).

[1] *Severe congestion due to oversubscription is a rare to extremely rare occurrence.*

[2] *Causes for Lost Credit are typically transmission errors such as ITW, CRC, or other signal related problems.*

6.1.8 Cisco

Cisco defines two different classes of congestion spreading events:

- Oversubscription

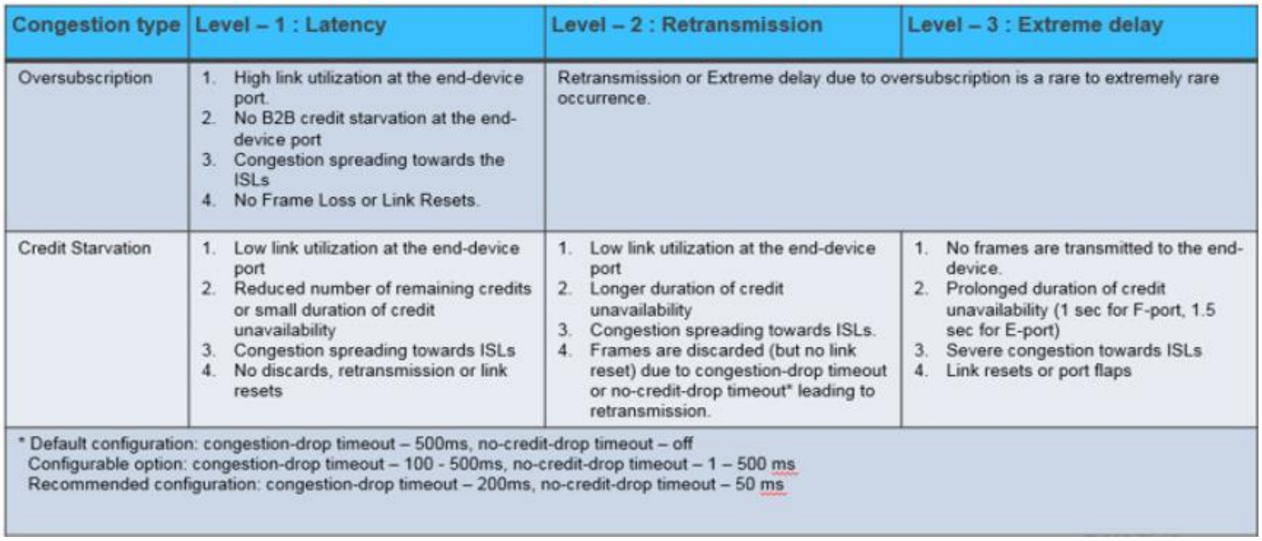
  As defined above.

- Credit Starvation

An end device or ISL that is not releasing credit at a rate fast enough to sustain line rate. For example, if an end device has negotiated a link speed of 16 Gbps and is unable to return credit at a rate that allows it to receive 16Gbps of data, then you could have a misbehaving device. These types of devices are also referred to as "slow drains". It's important to point out that a device might be misbehaving for a number of reasons including a driver problem or in the case of a ISL, because the port is experiencing the effects of congestion spreading.

- Cisco Congestion Spreading Terminology Cross Reference

Bringing all the pieces together for Cisco results in the following Cisco-specific Congestion Spreading Terminology cross reference.



Congestion type	Level – 1 : Latency	Level – 2 : Retransmission	Level – 3 : Extreme delay
Oversubscription	1. High link utilization at the end-device port. 2. No B2B credit starvation at the end-device port 3. Congestion spreading towards the ISLs 4. No Frame Loss or Link Resets.	Retransmission or Extreme delay due to oversubscription is a rare to extremely rare occurrence.	
Credit Starvation	1. Low link utilization at the end-device port 2. Reduced number of remaining credits or small duration of credit unavailability 3. Congestion spreading towards ISLs 4. No discards, retransmission or link resets	1. Low link utilization at the end-device port 2. Longer duration of credit unavailability 3. Congestion spreading towards ISLs. 4. Frames are discarded (but no link reset) due to congestion-drop timeout or no-credit-drop timeout* leading to retransmission.	1. No frames are transmitted to the end-device. 2. Prolonged duration of credit unavailability (1 sec for F-port, 1.5 sec for E-port) 3. Severe congestion towards ISLs 4. Link resets or port flaps

* Default configuration: congestion-drop timeout – 500ms, no-credit-drop timeout – off
Configurable option: congestion-drop timeout – 100 - 500ms, no-credit-drop timeout – 1 – 500 ms
Recommended configuration: congestion-drop timeout – 200ms, no-credit-drop timeout – 50 ms