

拥塞扩散以及如何避免此情况

本文档介绍了拥塞扩散（也称为慢漏）如何影响您的存储区域网络 (SAN)、用于描述 Connectrix B-Series 和 MDS Series 的各类拥塞严重性的指标，以及为避免拥塞扩散效应可以采取的预防性措施

2019 年 5 月

拥塞扩散以及如何避免此情况 | H17762.2 |

修订

日期	描述
2019 年 5 月	初始版本

答谢

本白皮书由 Dell EMC 存储工程团队的以下成员制作：

作者：
Alan Rajapa
Erik Smith

本出版物的内容按“原样”提供。Dell Inc. 对本出版物的内容不提供任何形式的陈述或担保，明确拒绝对适销性或针对特定目的的适用性进行默示担保。

使用、复制和发行本出版物所描述的任何软件都要有相应的软件许可。

© 2019 年 5 月发布：Dell Inc. 或其子公司。保留所有权利。Dell、EMC、Dell EMC 和其他商标为 Dell Inc. 或其子公司的商标。Dell® / EMC® / Dell EMC® 等品牌商标将有可能同时出现在戴尔易安信相关企业级产品（包括硬件和软件），和/或产品资料、戴尔易安信的官方网站。如果您有关戴尔易安信产品相关的任何疑问，欢迎联系您的指定客户经理。其他商标可能是其各自所有者的商标。

戴尔相信本文中所含信息在发布之日是正确的。该信息如有更改，恕不另行通知。

目录

- 1 前言 4
- 2 概述 5
 - 前提条件 5
- 3 什么是拥塞扩散? 7
- 4 因超额订阅而发生拥塞扩散 10
 - 应用程序基准 11
 - 生成应用程序基本配置文件图 11
 - 4.1.1 Brocade 17
 - 4.1.2 Cisco..... 18
 - UNISPHERE 拥塞扩散警报 19
 - 结论..... 24
- 5 补救 25
 - 预防措施 25
 - 带宽比率..... 25
 - 实施带宽限制 26
- 6 附录 28
 - 启用性能监视 28
 - CONNECTRIX 拥塞扩散监视 30
 - 6.1.1 Brocade 30
 - 6.1.2 Cisco..... 35
 - 6.1.3 Dell EMC 43
 - 6.1.4 Brocade 43
 - 6.1.5 Cisco..... 44
 - 6.1.6 超额订阅 45
 - 6.1.7 Brocade 46
 - 6.1.8 Cisco..... 48

本文档介绍了拥塞扩散（也称为慢漏）如何影响您的存储区域网络 (SAN)、用于描述 Connectrix B-Series 和 MDS Series 的各类拥塞严重性的指标，以及为避免拥塞扩散效应可以采取的预防性措施。

作为改善产品线性能和增强产品线功能的工作的一部分，Dell EMC 经常发布其硬件与软件产品的修订版。因此，对于当前使用的某些软件或硬件版本，本文档中介绍的一些功能可能不受支持。有关产品功能的最新信息，请参阅产品发行说明。

如果某产品不能正常运作或其功能与本文档的描述不符，请与您的 Dell EMC 代表联系。

受众

本技术书籍面向 Dell EMC 现场人员（包括技术顾问）以及参与了获取、管理、运营或设计包含 EMC 和主机设备的网络存储环境的存储架构师、管理员和操作员。

相关文档

所有相关文档和发行说明均可在 <https://dell.com/support> 找到。单击 **Support by Product**，输入产品名称并单击 **Documentation**。

Dell EMC 支持列表 和 E-Lab 导航 互操作性

如需获取最新信息，请始终参阅通过 E-Lab Interoperability Navigator (ELN) 提供的 *Dell EMC 支持列表*，网址为：

<https://www.dell.com/zh-cn/products/interoperability/elab.htm#tab0=2>

从何处获取帮助

Dell EMC 支持产品和许可信息可在 Dell EMC 在线支持网站上获取，如下所述。

注意：要通过 Dell EMC 在线支持网站创建服务请求，您必须具有有效的支持协议。请联系 Dell EMC 销售代表，以了解有关获取有效的支持协议的详细信息，或者获得有关您帐户的任何问题的回答。

产品信息

有关文档、发行说明、软件更新或有关 Dell EMC 产品、许可和服务的信息，请访问 Dell EMC 在线支持网站（需要注册），网址为：

<https://www.dell.com/support>

技术支持

Dell EMC 提供了多种多样的支持选项。

产品支持 —

Dell EMC 在 Web 上提供了特定于产品的整合信息，网址为：

<https://support.dell.com/products>

产品支持网页提供文档、白皮书、公告（例如常用的知识库文章）和下载的快速链接，以及更多动态内容，例如演示文稿、讨论、相关客户支持论坛条目和 Dell EMC 实时聊天链接。

Dell EMC 实时聊天 — 电子许可支持

打开与 Dell EMC 支持工程师的聊天或即时消息会话。

要激活您的授权并获得您的许可证文件，请按照用电子邮件发送给您的许可证授权代码 (LAC) 信函中的说明，访问 <https://dell.com/support> 上的 Service Center。

2

概述

本白皮书的目标是：

1. 描述拥塞扩散（也称为慢漏）如何影响您的存储区域网络 (SAN)，
2. 定义用于描述 B-Series 和 MDS Series 产品拥塞的各种严重性和类型的指标，
3. 描述可用于避免拥塞扩散效应的预防措施，以及
4. 演示如何使用上述信息来检测、防止和修复因超额订阅而发生的拥塞扩散。

前提条件

请注意：

本文档假定使用的是以下软件版本。在较旧版本中，这些步骤可能有所不同。

请参阅附录，了解有关如何启用所需功能的详细信息。

1. 已安装并正在运行 Dell EMC Unisphere for PowerMax 和 Unisphere for VMAX，并且已注册阵列来收集性能数据。
https://www.dell.com/support/products/27045_Unisphere-for-/Documentation/?source=promotion
2. 已安装 SAN 管理 GUI。
 - a. 对于 Brocade 结构：Connectrix Manager Data Center Edition (CMCNE) 14.x 或更高版本
下载：
https://www.dell.com/search/?text=CMCNE%2014&searchLang=en_US&facetResource=DOWN
管理指南：
https://www.dell.com/search/?text=CMCNE%2014%20admin%20guide&searchLang=en_US
 - b. 对于 Cisco 结构：Cisco Data Center Network Manager(DCNM) 10.x 或更高版本
下载：
<https://www.dell.com/support/search/?text=DCNM%2010&facetResource=DOWN>
管理指南：
<https://www.cisco.com/c/en/us/support/cloud-systems-management/prime-data-center-network-manager/products-installation-guides-list.html>
3. SAN 交换机固件应是以下各项：
 - a. Brocade：Fabric OS 7.4.1d 或更高版本
下载：
https://www.dell.com/support/search/?text=Brocade%20FOS%20download&searchLang=en_US&facetResource=DOWN
 - b. Cisco：NX-OS 6.2(13) 或更高版本
下载：
<https://www.dell.com/support/search/?text=NX-OS%20download>
4. 所有必需的性能监视许可证均已安装。
 - a. Brocade 需要 MAPS 许可证：

<https://docs.broadcom.com/docs/53-1005239-04>

b. Cisco 需要 DCNM-SAN 服务器产品包许可证：

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

c. PowerMAX 和 VMAX 需要 Dell EMC Unisphere 电子许可证。有关更多详细信息，请参阅以下 PDF 的第 21 页：

<https://www.dell.com/collateral/TechnicalDocument/docu88904.pdf>

什么是拥塞扩散？

3 什么是拥塞扩散？

在将数据传输到存储阵列以及从存储阵列传输数据时，都需要及时将数据传送到目的地。对于利用 SCSI 的基于数据块的存储协议（例如，光纤通道 — FCP）尤其如此。尽管这种情况的确切原因不在本白皮书的讨论范围之内，但《网络存储概念和协议》技术手册的“拥塞和背压”部分中提供了更多详细信息：

<https://www.dellemc.com/zh-cn/products/interoperability/elab.htm#tab0=1hardware/technical-documentation/h4331-networked-storage-cncpts-prtcls-sol-gde.pdf>。

与任何其他网络协议一样，光纤通道 (FC) 需要确保在各种常见网络拥塞情况下及时传输数据。FC 使用的机制重点关注通过使用缓冲区到缓冲区流控制来防止帧丢失。因此，FC 被视为“无损协议”。

尽管每个协议使用的流控制机制略有不同，但 FC 和其他无损协议（例如，DCB 以太网和 Infiniband）通过允许传送器确定链路另一端的接收器什么时候接近容量限制，来防止链路的任一端发生缓冲区溢出。确定此情况之后，一个端口将停止传输数据，直到链路的另一端指示它已准备好接收额外数据。当传送器处于此状态时，它无法传送帧，我们可以说它遇到拥塞。如果一个传送器出现拥塞的时间足够长，则此拥塞会向后传播到源。这种现象称为拥塞扩散，如以下一系列图表所示。

图 1 是未出现拥塞的 SAN 示例。主机 1 和主机 2 对阵列执行读取命令。由于阵列和主机均以 16 Gbps 速度连接，并且有足够的 ISL 带宽（即 32G），因此 SAN 中没有拥塞。

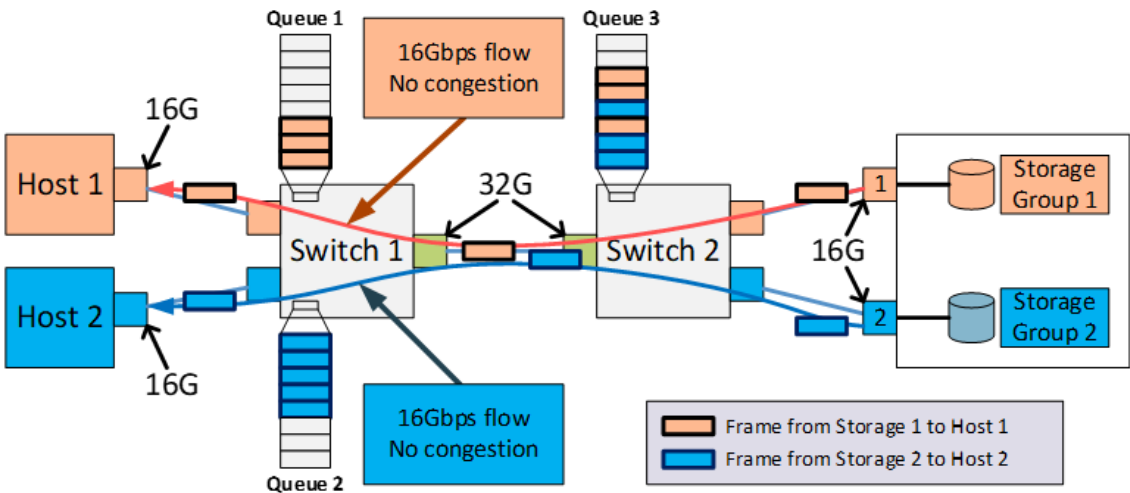


图1 无拥塞

图 2 显示了因超额订阅而发生拥塞扩散的 SAN 示例。请注意，两张图之间的唯一区别是，在图 3 中，主机 1 上的接口设置为以 4 Gbps 而不是 16 Gbps 运行。此操作完成后，如果阵列接口以大于已连接 HBA 的速度（即 4G）传输数据，则主机 1 将无法按传输速率接收数据，而且直接影响是造成帧排队。当队列 1 填满时，拥塞将扩散回数据的源。由于主机 1 和主机 2 共享相同的交换机间链路 (ISL)，所以此拥塞会影响主机 2 和存储 2 之间的“无辜流”，从而将吞吐量从 16 Gbps 降低为 4 Gbps。

拥塞扩散以及如何避免此情况[请在此处键入]

什么是拥塞扩散？

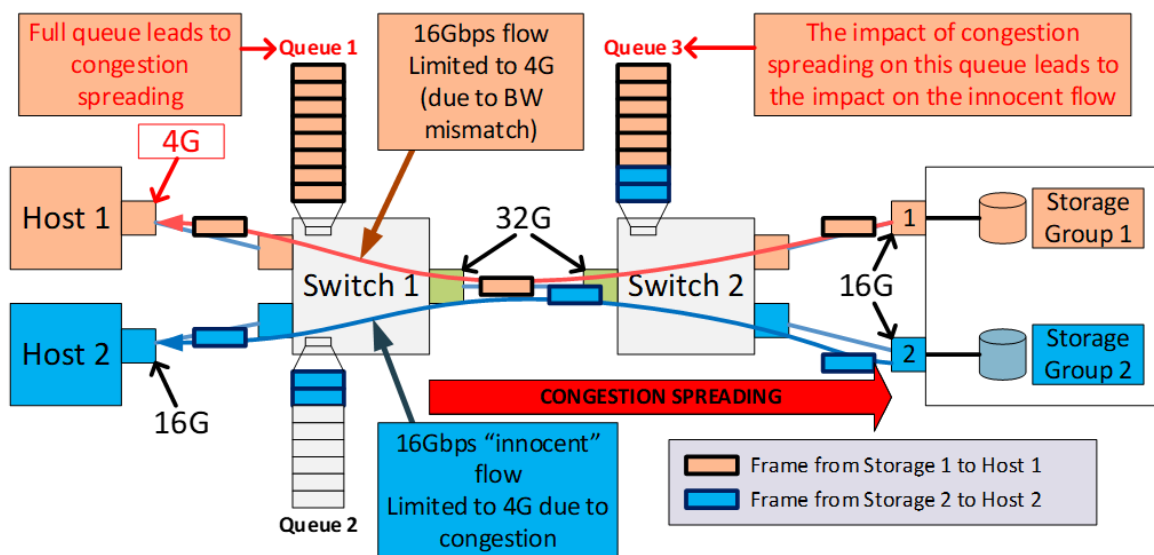


图2 拥塞

有关拥塞和拥塞扩散的其他信息可在《网络存储概念和协议》技术手册的“拥塞和背压”部分中找到：

(<https://www.dell.com/collateral/hardware/technical-documentation/h4331-networked-storage-cnpts-prtcls-sol-gde.pdf>)。

需要注意的是，超额订阅仅仅是拥塞扩散的潜在原因之一。其他原因将在以下各节中说明。

• 拥塞比率 (C Ratio)

拥塞比率或 C Ratio 是一个计算值，可帮助检测发生拥塞扩散的时间。例如，图 3 所示的主机（即主机 1）能够以 4 Gbps 的速率接收数据，但正在从能够以 16 Gbps 速率传输数据的存储接口接收数据。

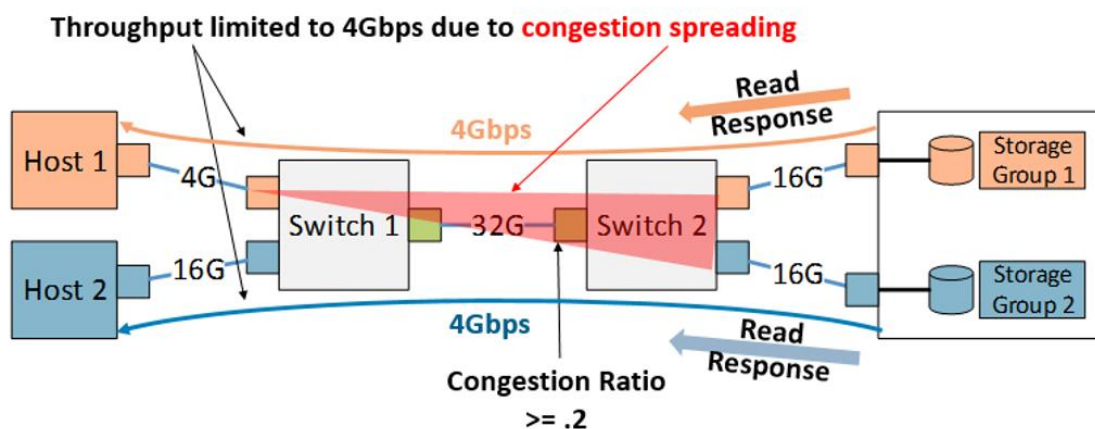


图3 拥塞比率

让这些问题很难检测和故障排除的部分原因是，从交换机 1 上的 4G 接口的角度来看，一切正常。交换机接口正在以链路允许的最快速度传输帧。但是，由于存储正在以其链路允许的速度（即 16 Gbps）传输数据，因此将会有 12 Gbps (16 Gbps - 4 Gbps) 的带宽由阵列用于传输，并且需要在某个地方排队。这种排队情况通常在结构 3 中发生，这是拥塞扩散的原

什么是拥塞扩散？

因。如前所述，可用于检测是否存在拥塞扩散的一种方法是计算拥塞比率。为此，可将“达到零传输信用所花的时间”计数器除以“已传输帧数”计数器，您将得到一个数字（通常介于 0 和 1 之间）。如果此数字大于 0.2，则表示出现拥塞。顺便说一下，此数字需要基于每个接口进行计算，因此最好为用于检查此值的过程编写脚本。

4 因超额订阅而发生拥塞扩散

以下案例分析基于因超额订阅造成的拥塞扩散。此案例分析的拓扑如下面的图 4 所示。在本案例分析中，您将了解当前可用于帮助检测并防止出现此问题的工具和技术。

注意：拥塞扩散是一个非常难以检测和解决的问题。这主要是因为无法通过当前一代管理工具清晰地指出问题正在发生，更不用说提供有关如何解决问题的任何实际指导了。因此，对这些问题进行故障排除需要终端用户了解问题是什么，然后了解如何使用当前可用的工具从有限的可用数据中得出结论。

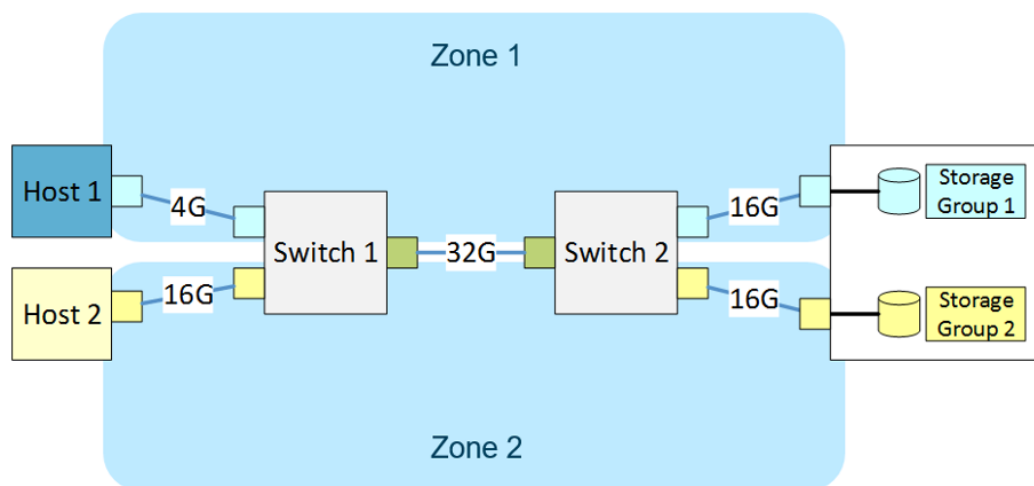


图 4 超额订阅案例分析的拓扑

- 情景：

用户 1 拥有在主机 2 (16G HBA) 上运行的现有应用程序，该应用程序以不同的数据块大小、队列深度和 I/O 模式运行 I/O。该应用程序在此环境中运行了很长时间，直到最近才遇到问题。本月早些时候，用户 2 决定在主机 1 (4G HBA) 上加载一个应用程序以进行测试。最初，在性能和延迟方面，环境中的一切都正常。但是，用户 1 最近发现其应用程序开始出现性能问题。

- 故障排除概述：

通常而言，要排查任何问题，您必须先了解事物在其理想情况下工作时的行为和配置。如您所知，SAN 有许多活动部件，这些部件共同构成生态系统，因此必须构建一个环境配置文件，其中包含构成 SAN 的三个主要组件的配置文件：应用程序、SAN 结构和存储。

在您环境中的各个组件上构建这些基准配置文件后，您将能够在出现问题时轻松地查明它们。应注意，这些配置文件不是一次性的。您应该在您环境的整个生命周期内不断收集基准数据，这样您不仅可以排查问题，还可为进一步增长和扩展进行规划。

在接下来的几个部分中，我们将介绍您必须从存储阵列收集这些基准统计信息，以便在上述情景下出现问题时，您已经做好充分准备，可以找到根本原因。

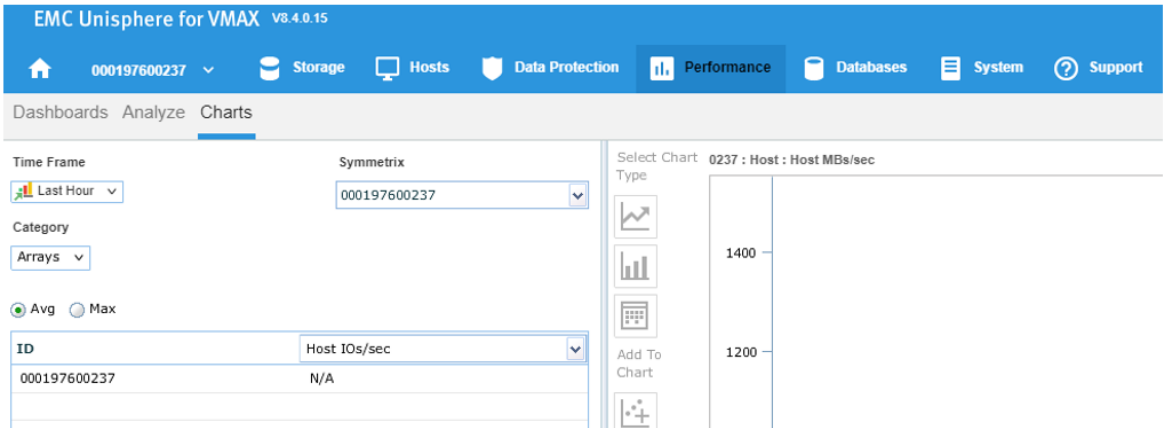
因超额订阅而发生拥塞扩散

应用程序基准

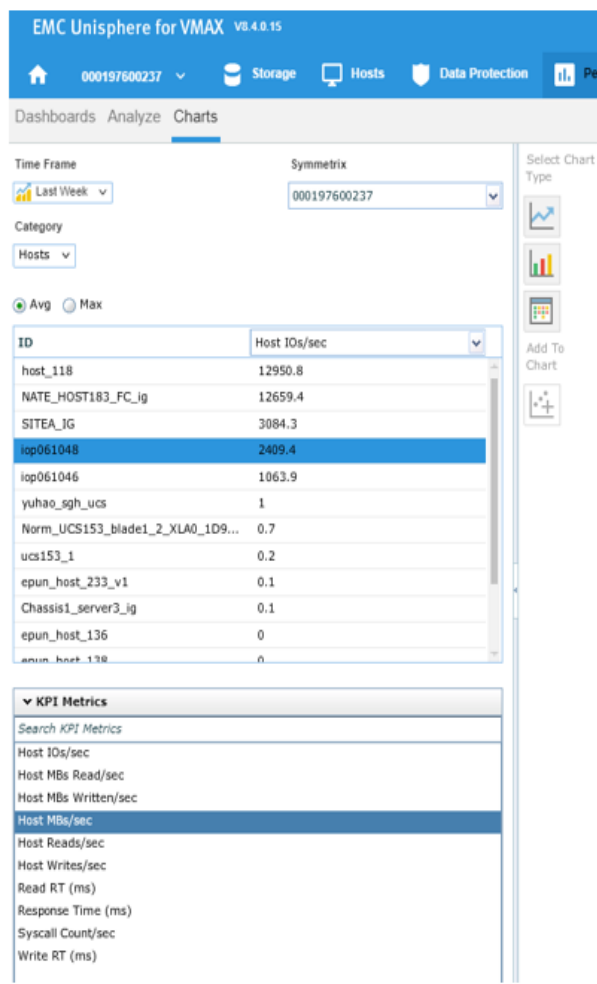
在使用 Dell EMC PowerMax 和 VMAX 的情况下，如果您启用性能监视，则可以返回历史记录（自启用此功能之时起最长一年），以便了解在出现任何变化之前，您的应用程序配置文件在 IOPs 和响应时间方面是何种情况。拥有此应用程序基本配置文件让您可以使用生成的图表，轻松确定可能存在问题的位置。

生成应用程序基本配置文件图

在 Dell EMC Unisphere 中，单击 **Performance > Charts**



选择 **Time Frame**。这是在您注意到性能问题之前的任何时间。在 **Category** 下拉菜单中，选择 **Hosts > Hosts**。



选择有问题的主机。对于 **KPI Metrics**，我们将生成七个不同的图表。针对每个 KPI 指标重复此部分。如果您一次单击所有指标，则图表会将它们放在单个图形中。

- a. Host I/Os/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)

在图 5 “主机每秒 IO 和 MB” 中，我们要查看主机每秒 IO 和 MB 数。在这些图表中，我们可以看到应用程序什么时候运行最多 IO 并利用链路的全部带宽，以及什么时候运行最少 IO，并且还可以看到它们分别持续多长时间。

注意：在图例中，您会发现共有两个主机，但当前仅显示其中一个主机的 IO，因为另一个主机未执行任何 IO。



图5 每秒主机 IO 和 MB

图 6 “每秒读取和写入数”中显示的图表提供了应用程序生成的 IO 类型的细分。根据这些图表，我们可以确定应用程序 IO 中读取和写入所占的百分比。在本例中，我们可以确认应用程序的读/写比大约为 70/30。

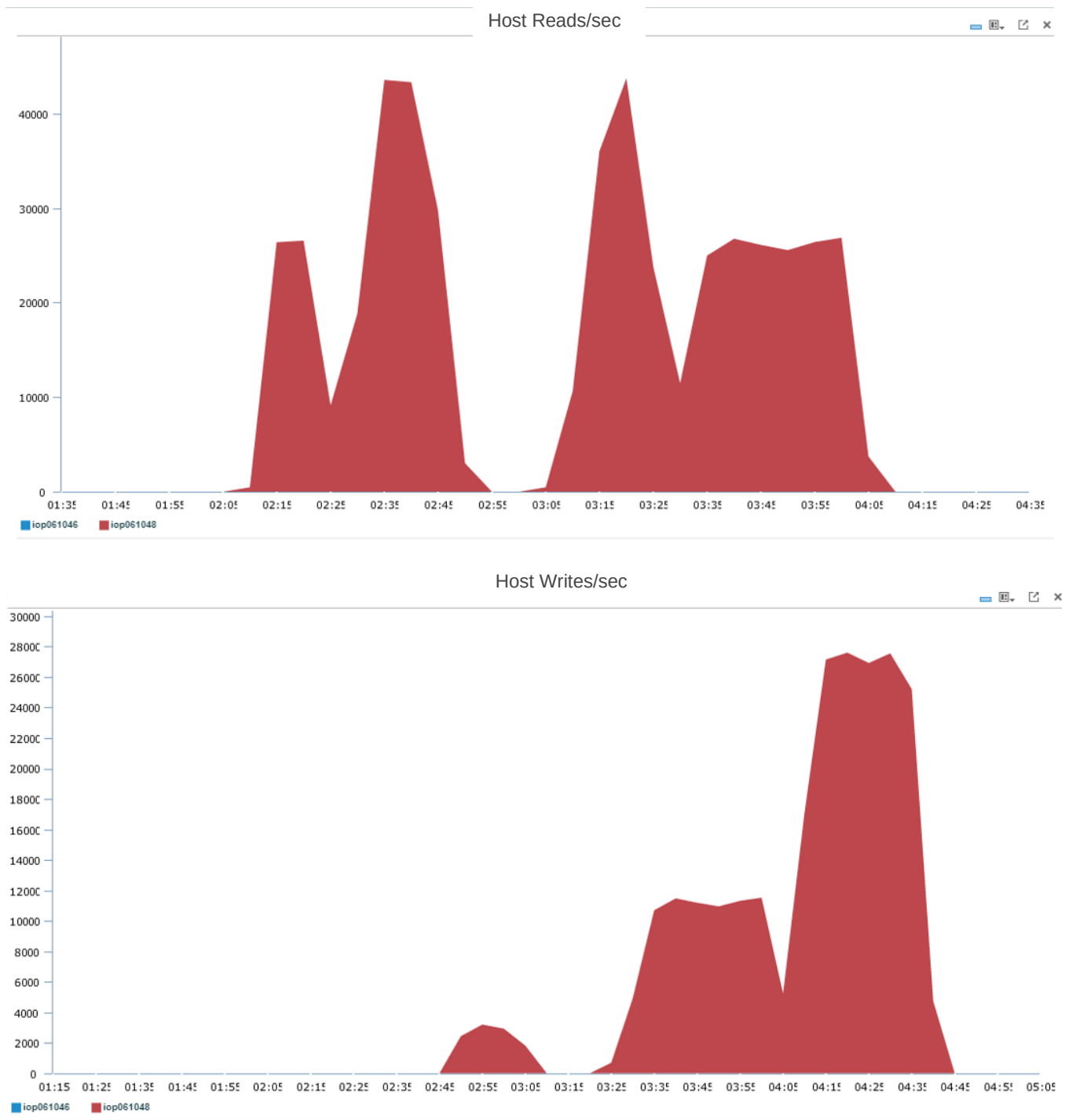


图6 每秒读取和写入数

图 7 和图 8 可能是在进行故障排除时最有用的图表。它们提供读取和写入的响应时间细分，从而使我们能够了解该应用程序遇到的延迟。当我们需要对性能问题进行故障排除时，这非常有用，因为如果响应时间存在峰值，则我们可以使用之前的图表将该峰值与特定事件关联起来。

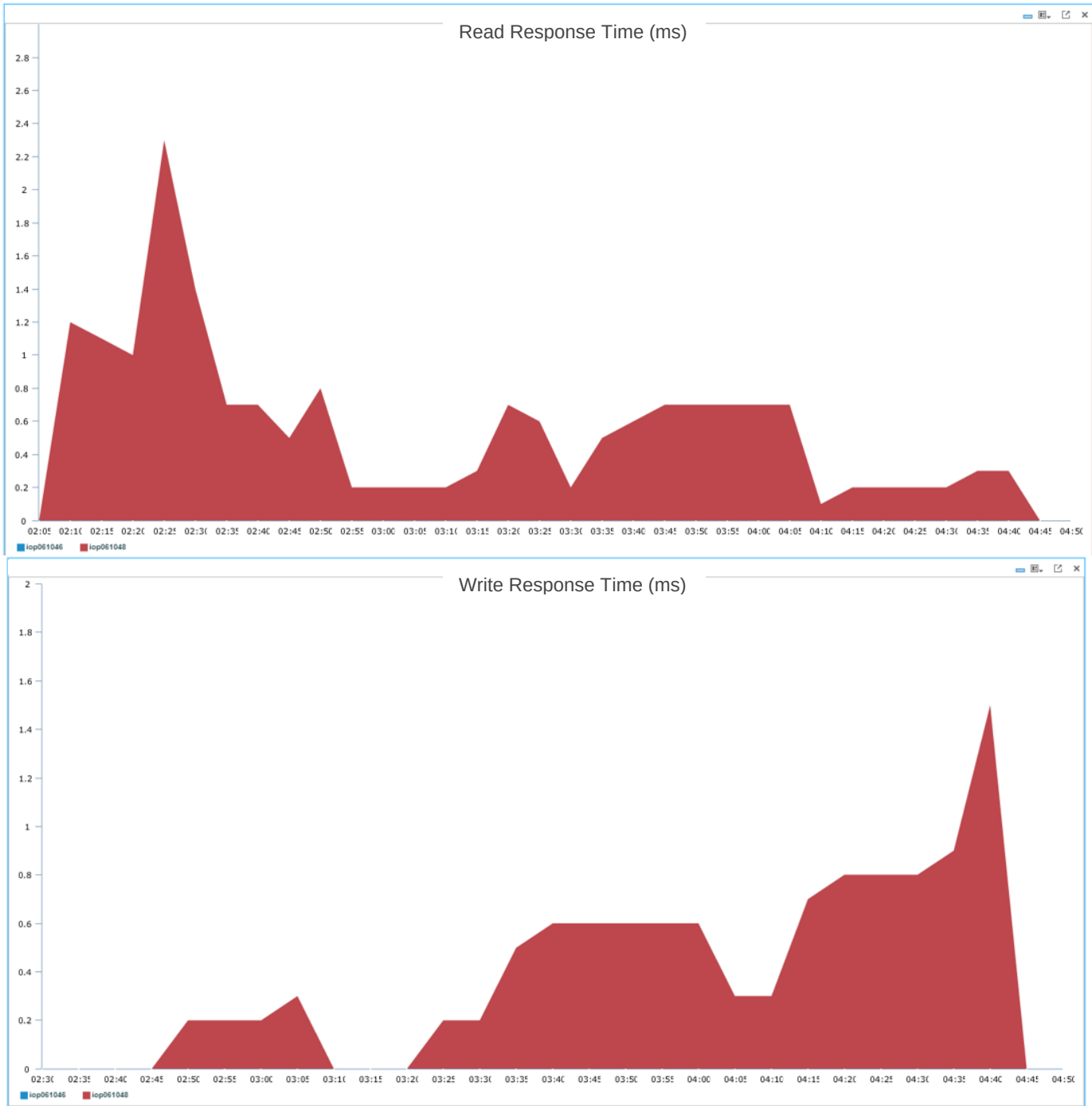


图7 读取和写入响应时间 (毫秒)

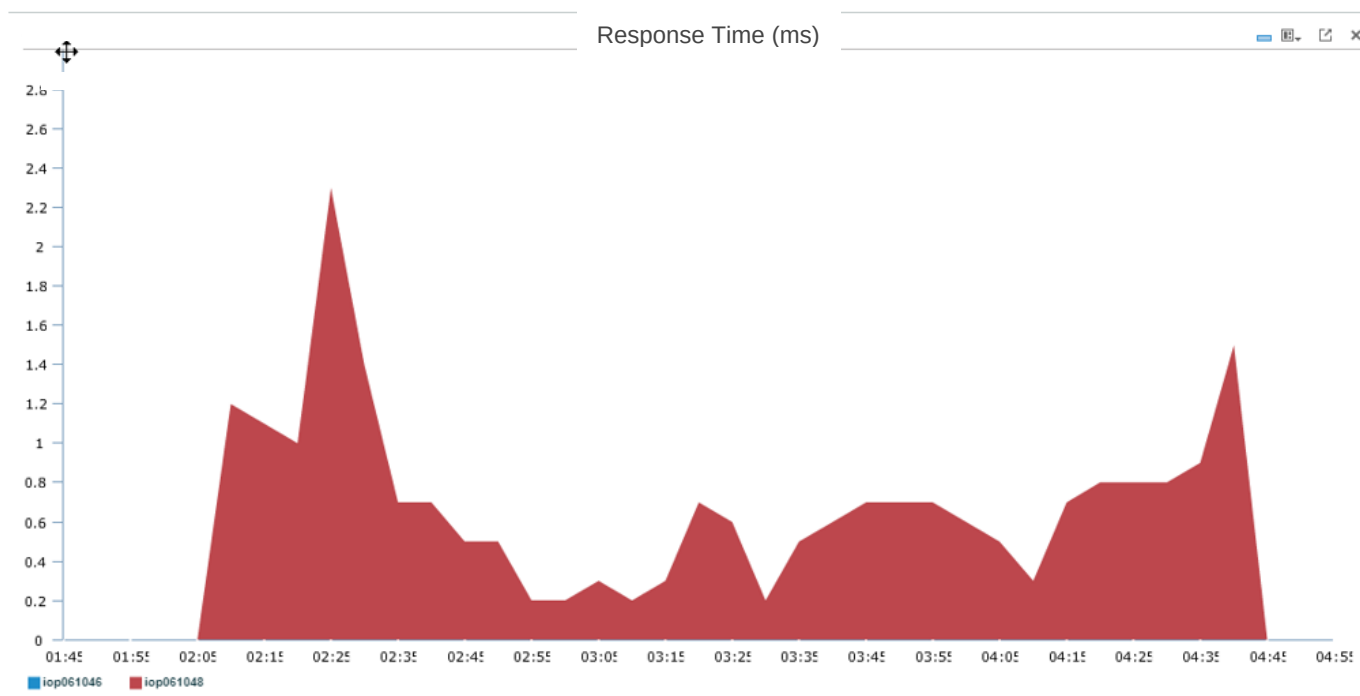


图8 响应时间 (毫秒)

现在，我们有用户 1 应用程序的配置文件；我们知道其读取/写入比大约为 70/30，平均响应时间大约为 0.7ms，最长响应时间为 2.3ms。

正如[情景部分](#)中讨论的那样，我们最近添加了一个新主机，环境中开始出现性能问题，因此让我们来看看如何排查该问题。

由于我们的环境中出现性能问题，因此我们需要实施 SAN 中提供的功能，帮助我们确定何时出现这些类型的问题。

因为我们了解我们的平均响应时间是多长，所以根据应用程序配置文件，我们知道此性能问题是由于高于预期响应时间导致的。

Connectrix SAN 拥塞扩散警报

在本节中，我们将查看我们从 SAN 交换机侧报告的拥塞事件的类型。请确保您已按照[前提条件](#)中所述在环境中启用了这些功能。

4.1.1 Brocade

- 1. 确保您的控制面板上至少有 **Top Port Traffic** 和 **BB Credit Zero**。如果还未添加它们，请单击左上角的扳手进行添加。

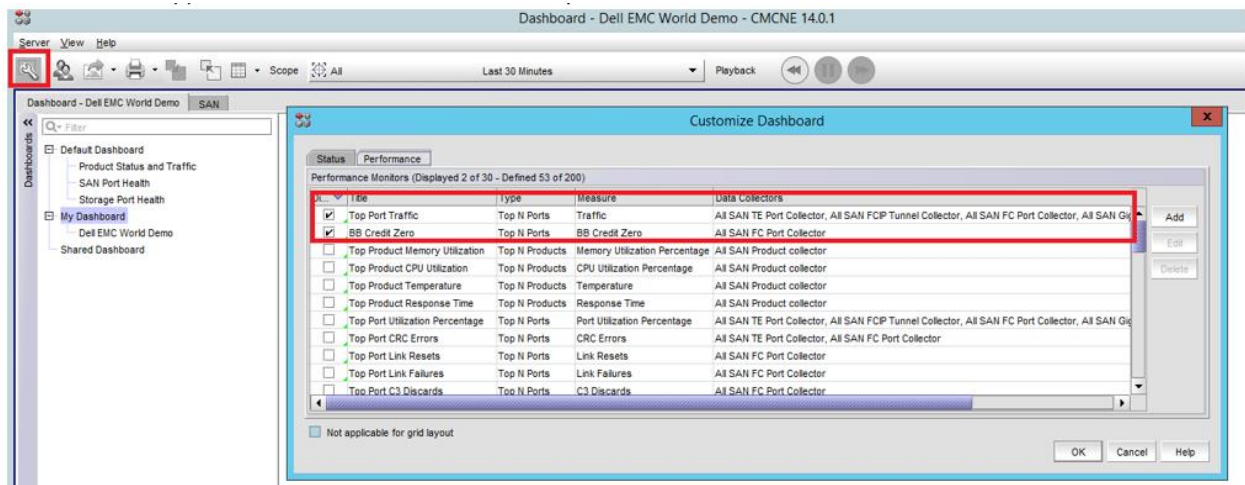


图9 CMCNE 控制面板

- 2. 因超额订阅而发生拥塞扩散时，正如图 9 “CMCNE 控制面板” 中所示，您通常会在 CMCNE 控制面板中看到以下警报：

- a. Highly utilized F-Port
- b. BB Credit Zero

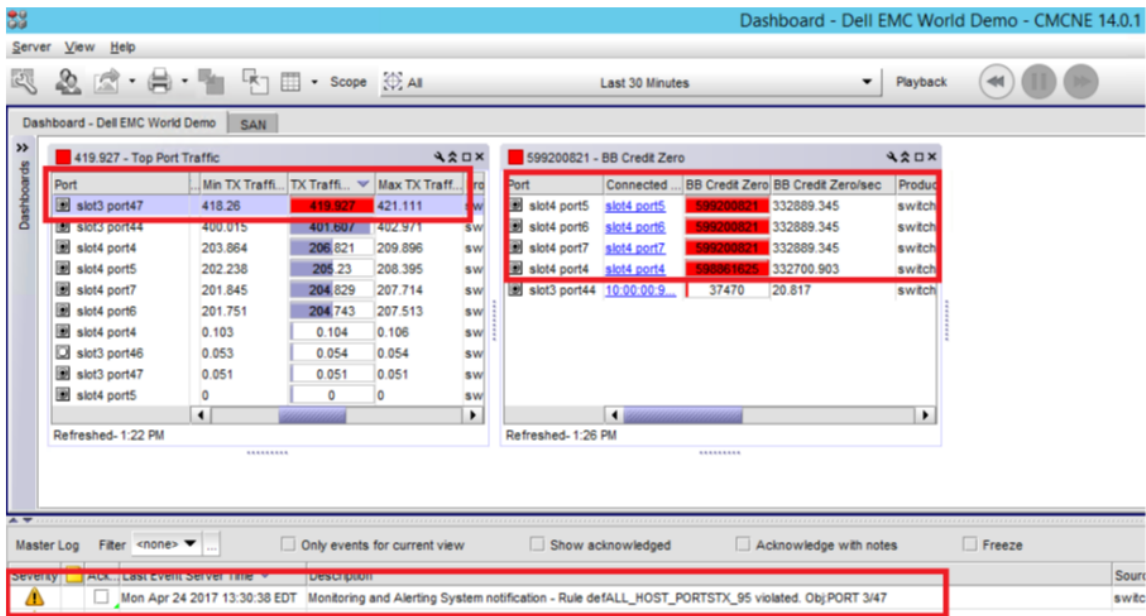
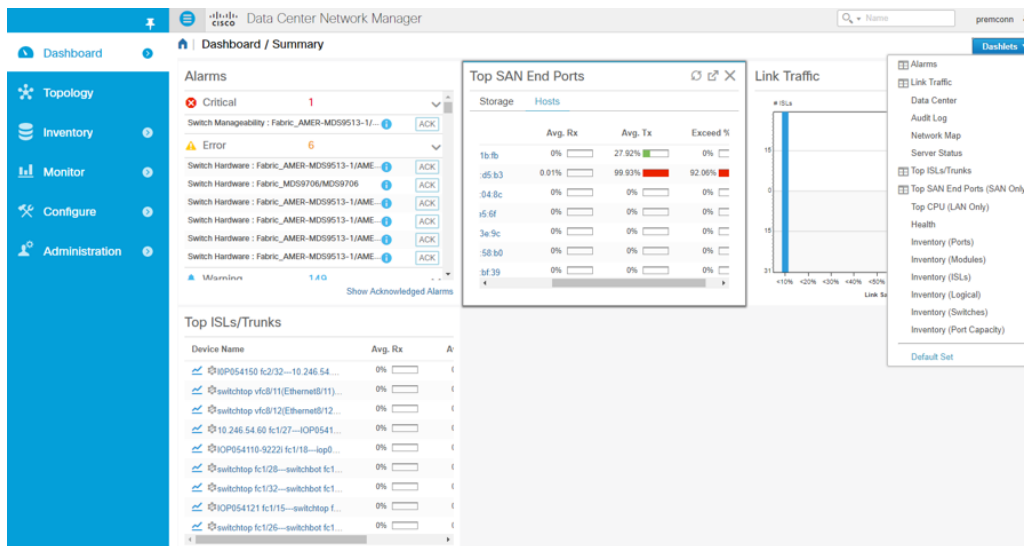


图10 CMCNE 控制面板显示警报

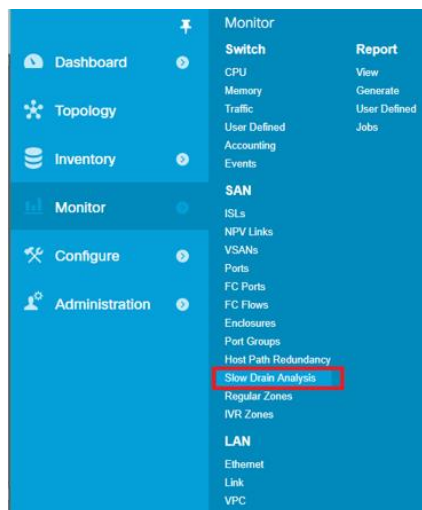
- 这两个事件的组合（ISL 上出现使用率极高的 F 端口和高缓冲区到缓冲区信用变为零）可能意味着出现需要进行调查的潜在性能问题。请参阅“补救”一章，了解有关检查内容的步骤。

4.1.2 Cisco

- 在 DCNM 控制面板中，您应该有 **Top SAN End Ports** 控制面板小程序。如果没有，您可以从下拉菜单中添加它。在 **Top SAN End Ports** 中，您将看到一个设备报告其利用率超过 90%。DCNM 有默认阈值，这将导致在开始超过默认利用率时它会将端口标记为黄色或红色。此警报本身并不一定意味着 SAN 中存在性能问题。我们还需要寻找结构中的其他警报。



- 如果您看到使用率过高的 F 端口，请运行慢漏分析工具。
单击 **Monitor > SAN > Slow Drain Analysis**。



- 运行慢漏分析工具 10 分钟。报告完成后，您会注意到，在报告运行的时段内，大量 TxWait 计数器都有递增。这些警报与 F 端口使用率高警报相结合表示由于超额订阅而发生了 SAN 拥塞。

因超额订阅而发生拥塞扩散

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

Slow Drain Details for undefined

Zoom: 10 minutesMAX

Total 13

Show Non Zero data rows only

Interface	Speed	ConnectTo	Type	Level 3			Level 2		Level 1		
				TxCreditLoss	TxLinkReset	RxLinkRe...	TxTimeoutD...	TxDiscard	TxWtAvg10...	RxB2Bto0	TxB2Bto0
fc2/36 ...	16Gb	ISP054151 fc2/36 (port-channel6)	Switch	0	0	0	0	0	0	0	41429192
fc2/37 ...	16Gb	ISP054151 fc2/37 (port-channel6)	Switch	0	0	0	0	0	0	0	30150028
fc2/38 ...	16Gb	ISP054151 fc2/38 (port-channel6)	Switch	0	0	0	0	0	0	0	26941217

这两个事件的组合（ISL 上出现使用率极高的 F 端口和高缓冲区到缓冲区信用变为零）可能意味着出现应该进行调查的潜在性能问题。请参阅[补救](#)部分，了解有关检查内容的步骤。

UNISPHERE 拥塞扩散警报

在本节中，我们将了解如何使用 Unisphere for PowerMAX 和 VMAX 将 SAN 交换机事件关联到存储阵列。

请确保您已按照“前提条件”部分中所述在环境中启用了这些功能。

- 1. 按照[生成应用程序基本配置文件图](#)中的步骤操作。生成同样的七个图表，将用户 2 的主机添加到组合中（因为这是在出现性能问题之前环境中最近发生的一项更改）。审核数据。

请记住，我们的应用程序基本配置文件 = 70/30 读取/写入比率，平均响应时间大约为 0.7ms，最长响应时间为 2.3ms。

在下面的图 11 中，我们比较了每秒 IO 和 MB 数，我们实际上并没看到有问题的迹象。事实上，如果您将其与原始基准应用程序图表进行比较，会发现我们实际上完成了更多的 IOPs。

此外，您可以看到在有些点，我们接近线速（下面突出显示）。这些点将在以后发挥作用。

因超额订阅而发生拥塞扩散



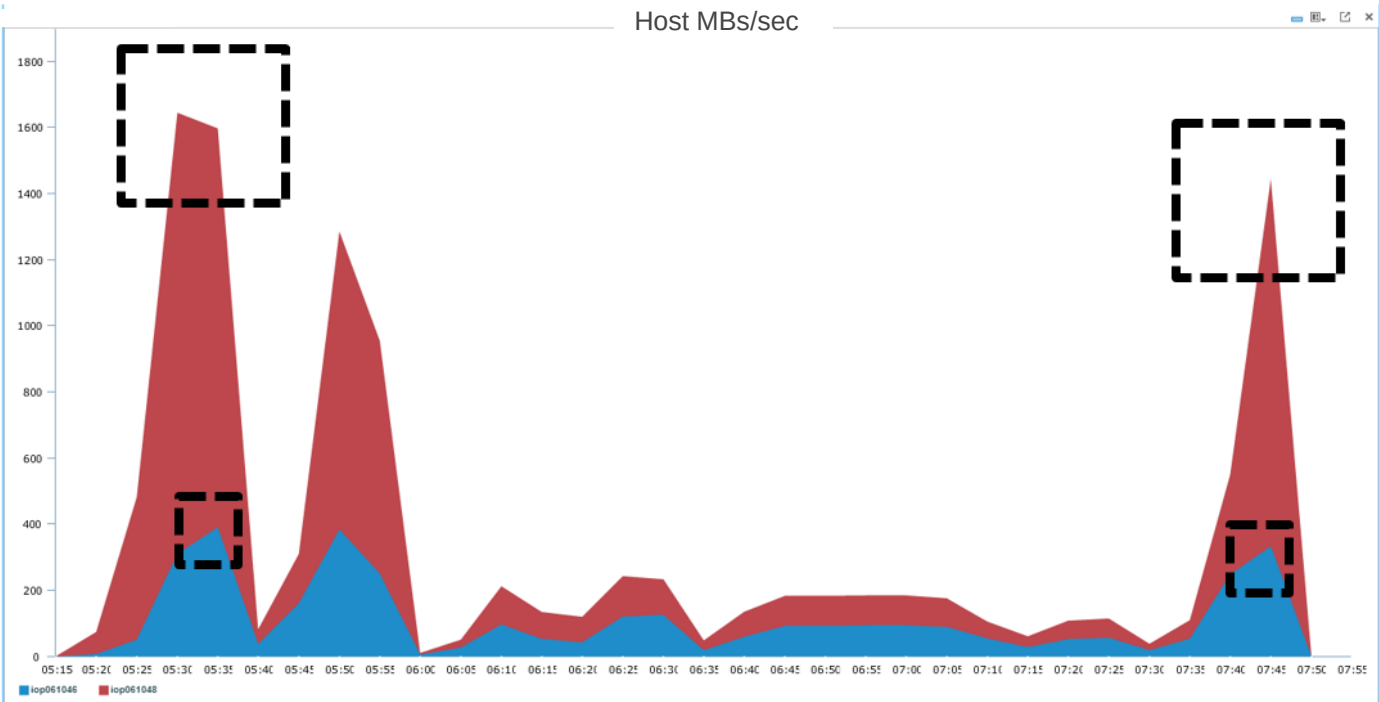


图11 每秒主机IO 和MB

图 12 显示两台服务器之间的读取/写入比的比较，您可以看到此时两台服务器之间的 IO 情况没有很大差异。

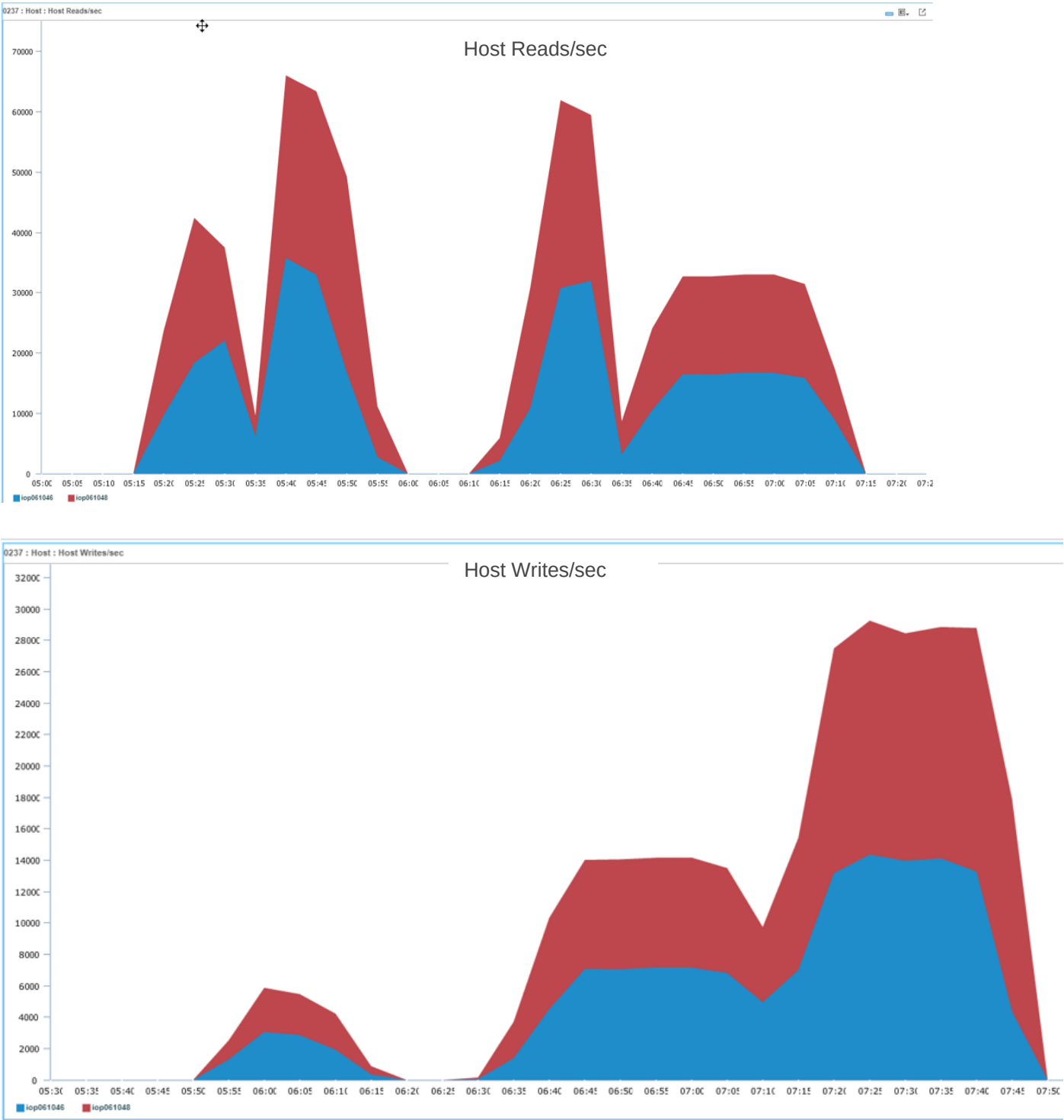
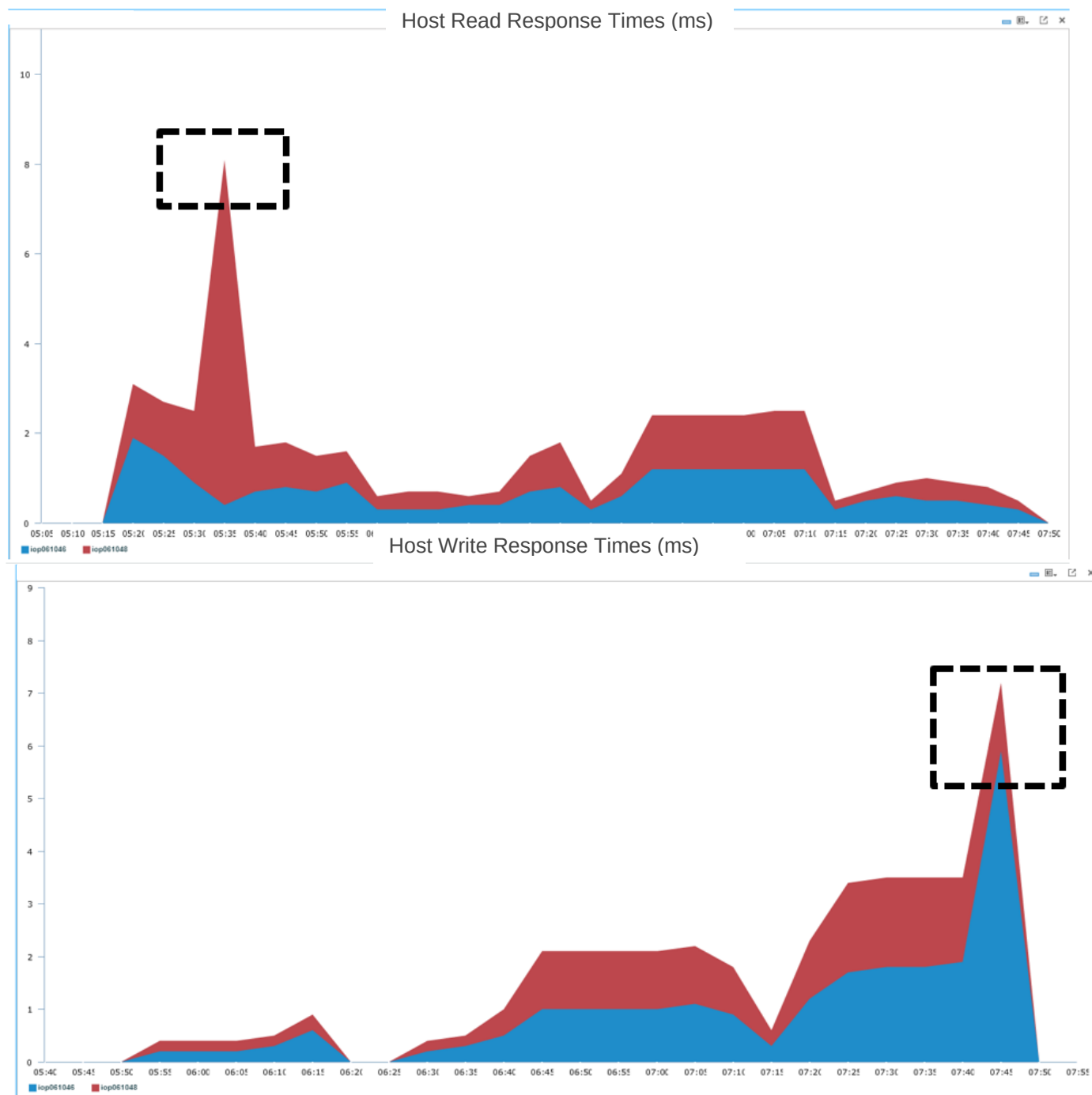


图 12 每秒主机读取和写入数

图 13 提供了最有用的信息。记住我们的应用程序配置文件，我们看到的平均响应时间大约为 0.7ms，最长响应时间为 2.3ms。从下面的图表中，我们可以看到，当我们进入 8ms 范围时，响应时间出现突增的高峰，并且我们的总体平均响应时间也在增加。

回头看一下图 11，我们可以看到，这些高响应时间与两台服务器接近线速的时间相关联。通常情况下，在光纤通道中，您需要使用大数据块 IO（大于 128k）才能使链路达到饱和。



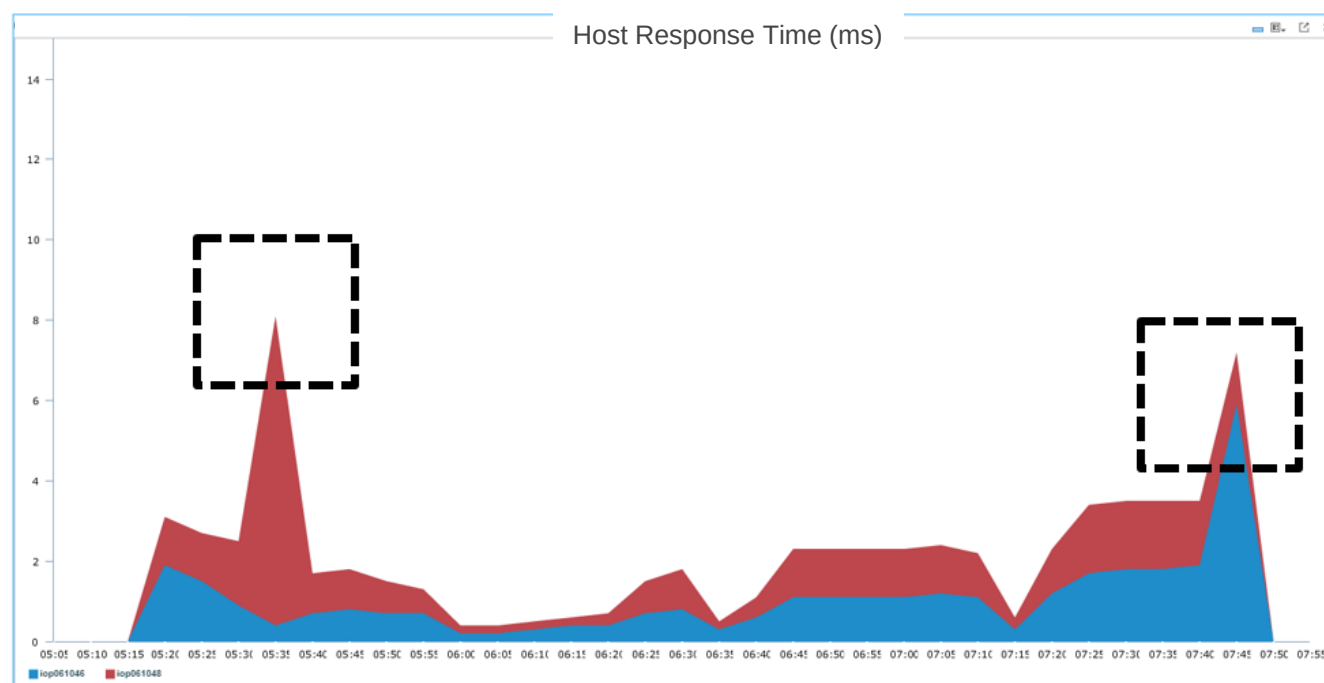


图13 主机读取和写入响应时间（毫秒）

结论

总结一下我们到目前为止在此案例分析中了解的所有信息：

- **Connectrix SAN：**

1. 我们的 SAN 报告有大量的缓冲区到缓冲区信用变为零。
2. 我们在 F 端口上看到高流量利用率。

- **Dell EMC PowerMAX 和 VMAX：**

1. 完全链路利用期间出现长响应时间

如前所述，由于带宽不匹配造成的拥塞非常难以检测并使用如今可用的工具集进行确认。但是，根据以上警报，我们可以推断出该问题是由带宽不匹配和大数据块读/写造成的。这可通过完全链路利用期间出现长响应时间来指示。

检测此问题的另一种方法是使用**拥塞比率**。如今，我们必须在环境中手动计算此项（您也可以尝试为其编写脚本），但我們知道一旦您的 C-Ratio 大于 0.2，因为 SAN 环境中出现背压，所以您会遇到拥塞。C-Ratio 将是出现慢漏的第一个指示。

5 补救

预防措施

对于此特定案例分析（因超额订阅而造成拥塞扩散），您可以在环境中部署几个选项，帮助防止出现此问题。

带宽比率

- 在检查 SAN 时，您想要确定以较低速度运行的设备，然后了解它们的应用程序流量特点的类型。请记住，只是因为出现带宽不匹配并不意味着一定存在问题。
- 端到端检查结构以确保所有终端设备都以相同的链路速度运行。
- 确保 ISL 上有充足的带宽。一个良好的经验法则是，总 ISL 带宽应尽可能等于或大于结构中的总存储带宽量。
- 您可以通过确保升级所有端到端组件（如下面的图 14 所示），让整个 SAN 实现现代化。这种方法的缺点是，来自端到端的零超额订阅在大型环境中不现实。此外，它的成本会非常高昂。因此，您还应该仅将重点放在升级特定主机、交换机和存储连接上。

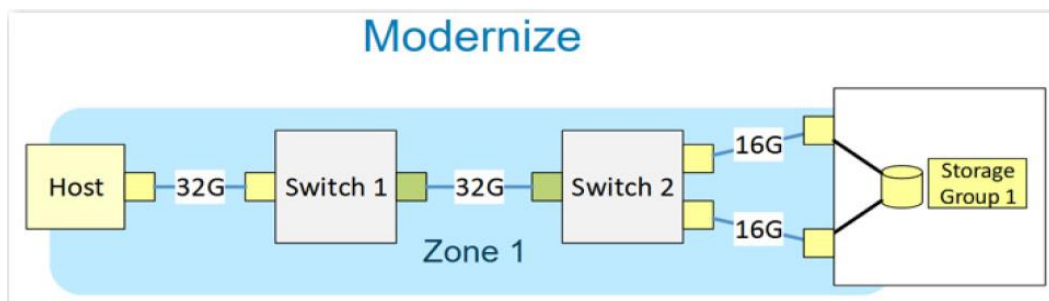


图 14 现代化

另一种方法是重新分区，如图 15 所示。

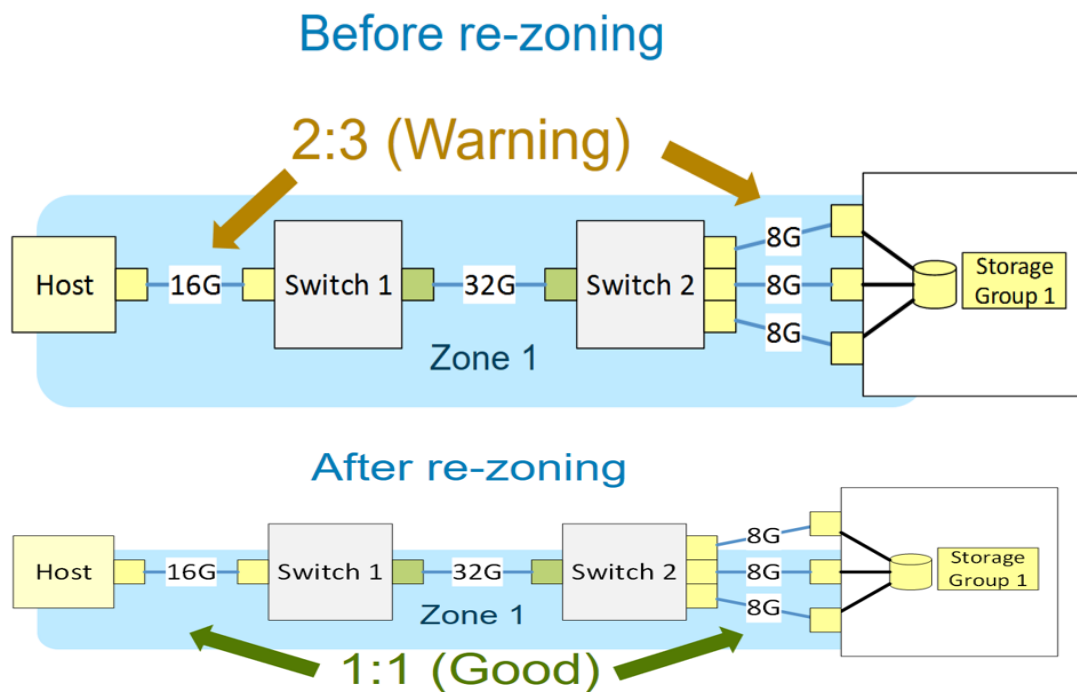


图 15 重新分区之前和之后

实施带宽限制

Dell EMC VMAX 和 Dell EMC Unity 平台对存储组 (VMAX) 或 LUN (Unity) 施加带宽限制。在上面的案例分析中，由于超额订阅导致拥塞扩散，当我们实施带宽限制时，我们看到性能恢复，如下面的图 16 所示。这可以在存储组上直接通过 Unisphere 完成。

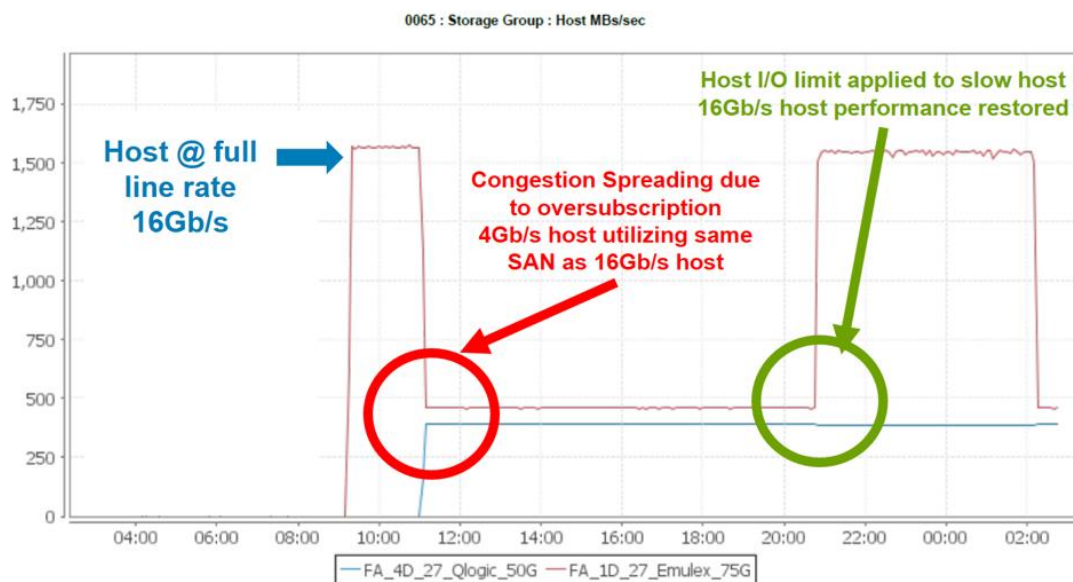


图 16 已应用主机 I/O 限制

对于 I/O 限制，必须注意，这并不适用于群集。例如，让我们来看下面的图 17。当主机限制应用于产生背压的 4 Gb 主机时，阵列开始限制其发送回 4 Gb 主机的数据量（基于 IO 限制设置），因此您可以完全消除背压问题，而其他流可以按全线速运行。

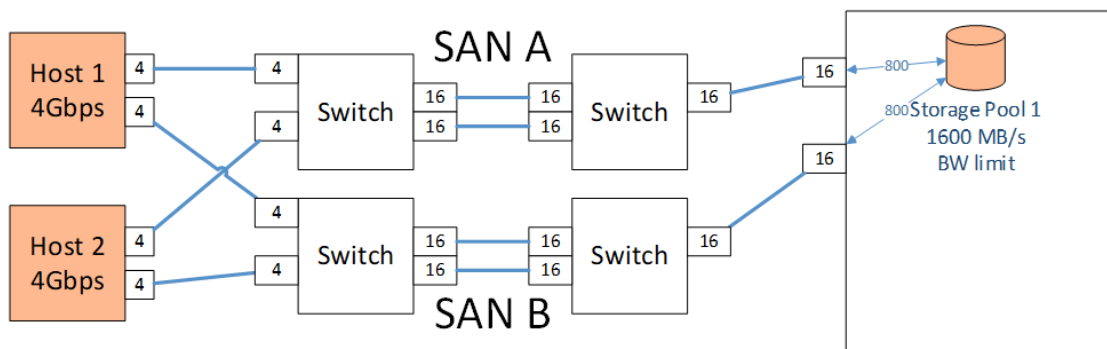


图 17 群集的 I/O 限制

在此示例中，我们在群集中有两个以 4 Gb/s 速率运行的主机。因为它们在群集中，这两个主机将通过每个结构访问卷，因此我们需要将带宽 I/O 限制设置为 1600 MB/s（每个 FA 为 800 MB/s）。但是，使用这种方法，没有什么能阻止单个 HBA 使用全部 800 MB/s。

- 隔离

防止出现此问题的另一种方法是将速度较慢的流量与高速流量隔离开来，并使用专用 ISL。这可以通过创建虚拟结构 (Brocade) 或 VSAN (Cisco) 来实现，

如下面的图 18 所示。这种方法的缺点是，您必须使用专用端口，但这会使较慢的流量不会影响您的更高速度流量。要在 Brocade 上启用虚拟结构，它将需要停机时间，因为整个交换机都必须重新启动。将端口移到 Cisco 上的不同 VSAN 时，只会影响要移动的终端设备。

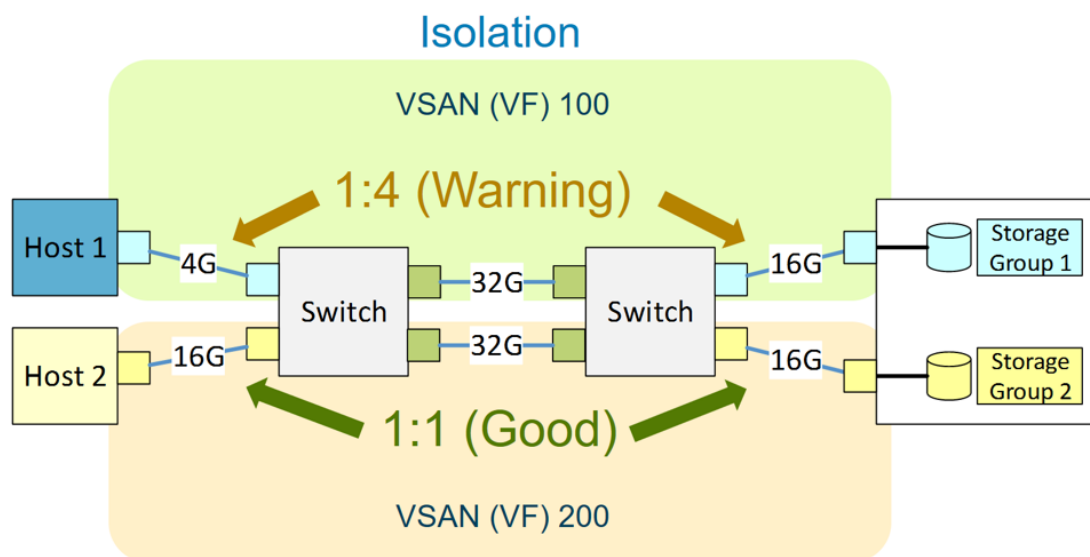


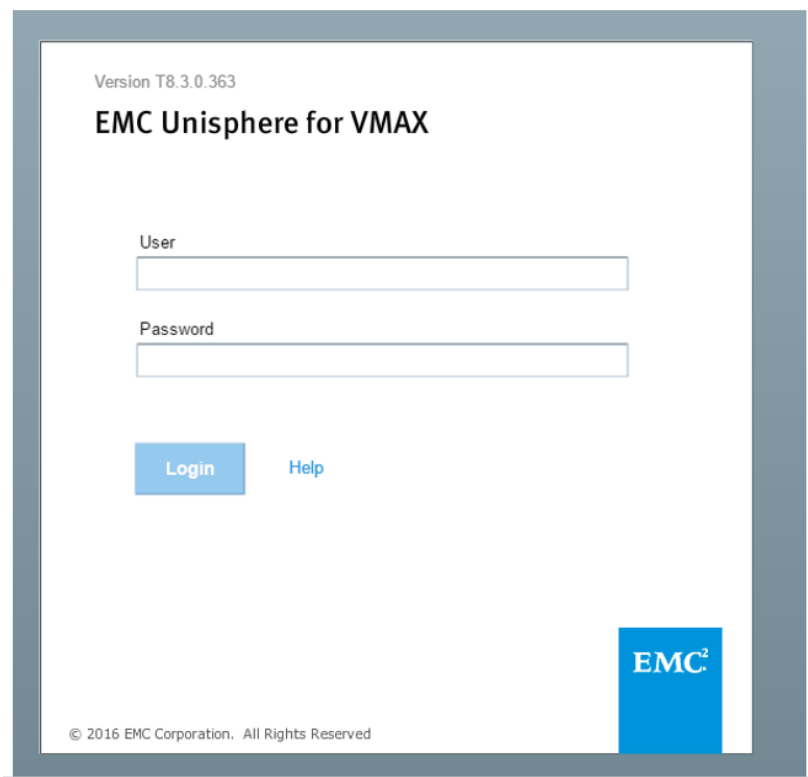
图 18 隔离

6 附录

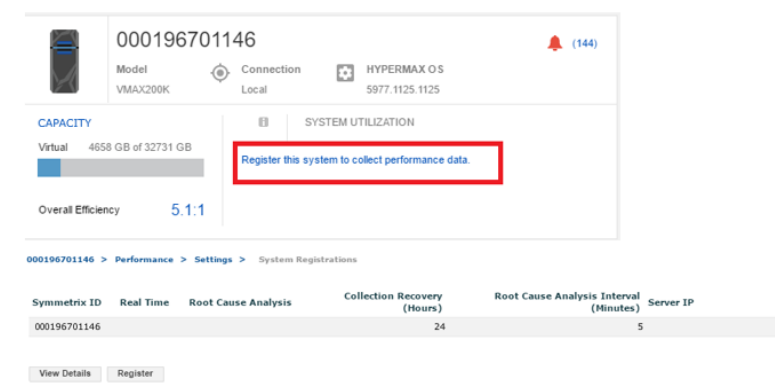
启用性能监视

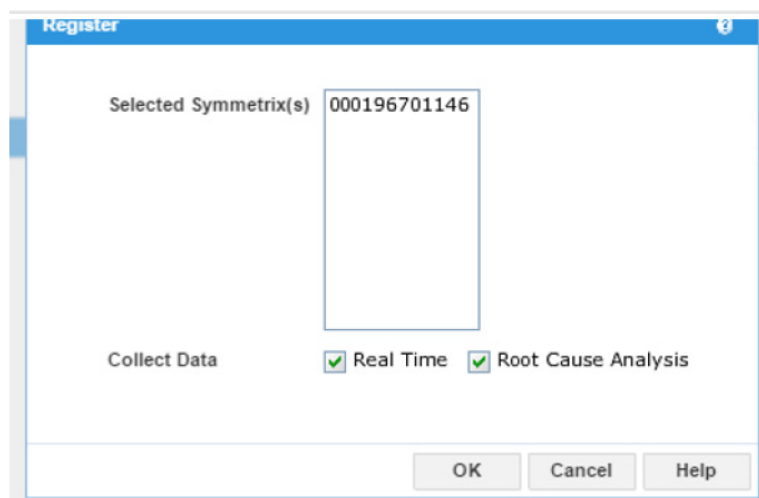
本节介绍有关如何在 Unisphere for VMAX 中启用和审查性能监视数据的步骤。

- 1. 登录 Unisphere GUI。



- 2. 确保阵列已注册以收集性能数据。如果未注册，请注册阵列。





The image shows a 'Register' dialog box with a blue title bar. Inside, there is a label 'Selected Symmetrix(s)' followed by a text box containing the value '000196701146'. Below this, the text 'Collect Data' is followed by two checked checkboxes: 'Real Time' and 'Root Cause Analysis'. At the bottom right, there are three buttons: 'OK', 'Cancel', and 'Help'.

Register

Selected Symmetrix(s) 000196701146

Collect Data ☒ Real Time ☒ Root Cause Analysis

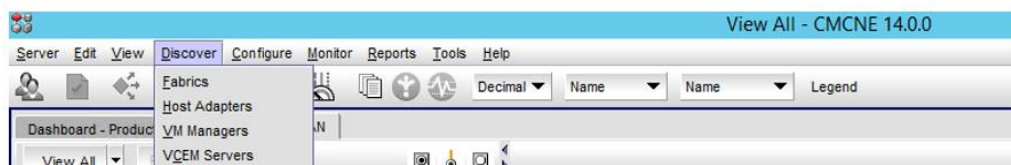
OK Cancel Help

CONNECTRIX 拥塞扩散监视

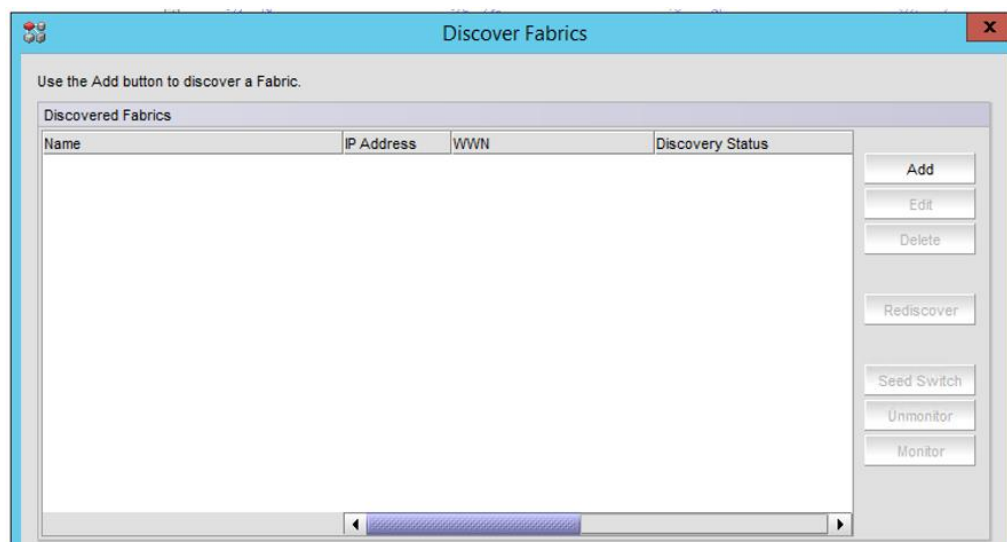
6.1.1 Brocade

- 发现结构

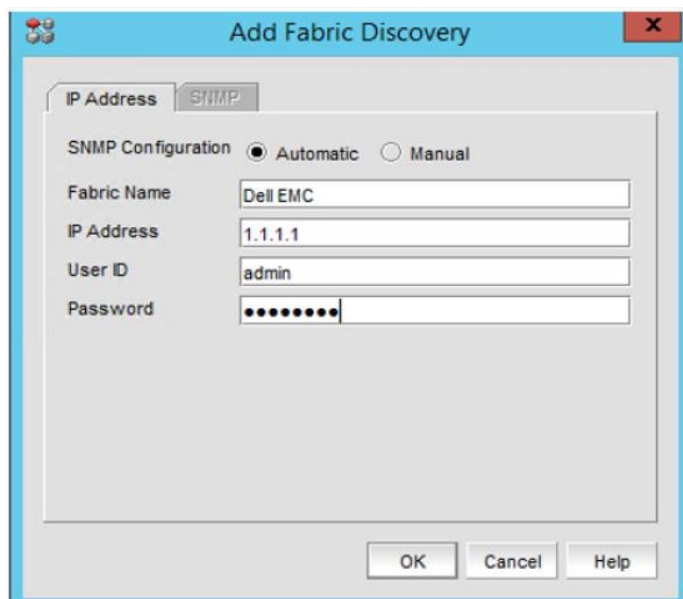
1. 登录到 CMCNE 服务器，然后单击 Discover > Fabrics (SANnav?)



- 2 在新窗口中，单击 **Add**。



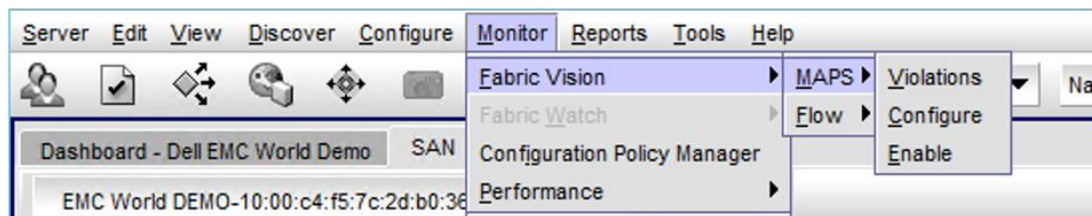
- 3 为结构中的其中一个交换机填写所需的信息。CMCNE 将自动发现该结构中的所有交换机，并假定结构中的所有交换机的用户名和密码都相同。



4 对所有其他结构重复此部分的操作。

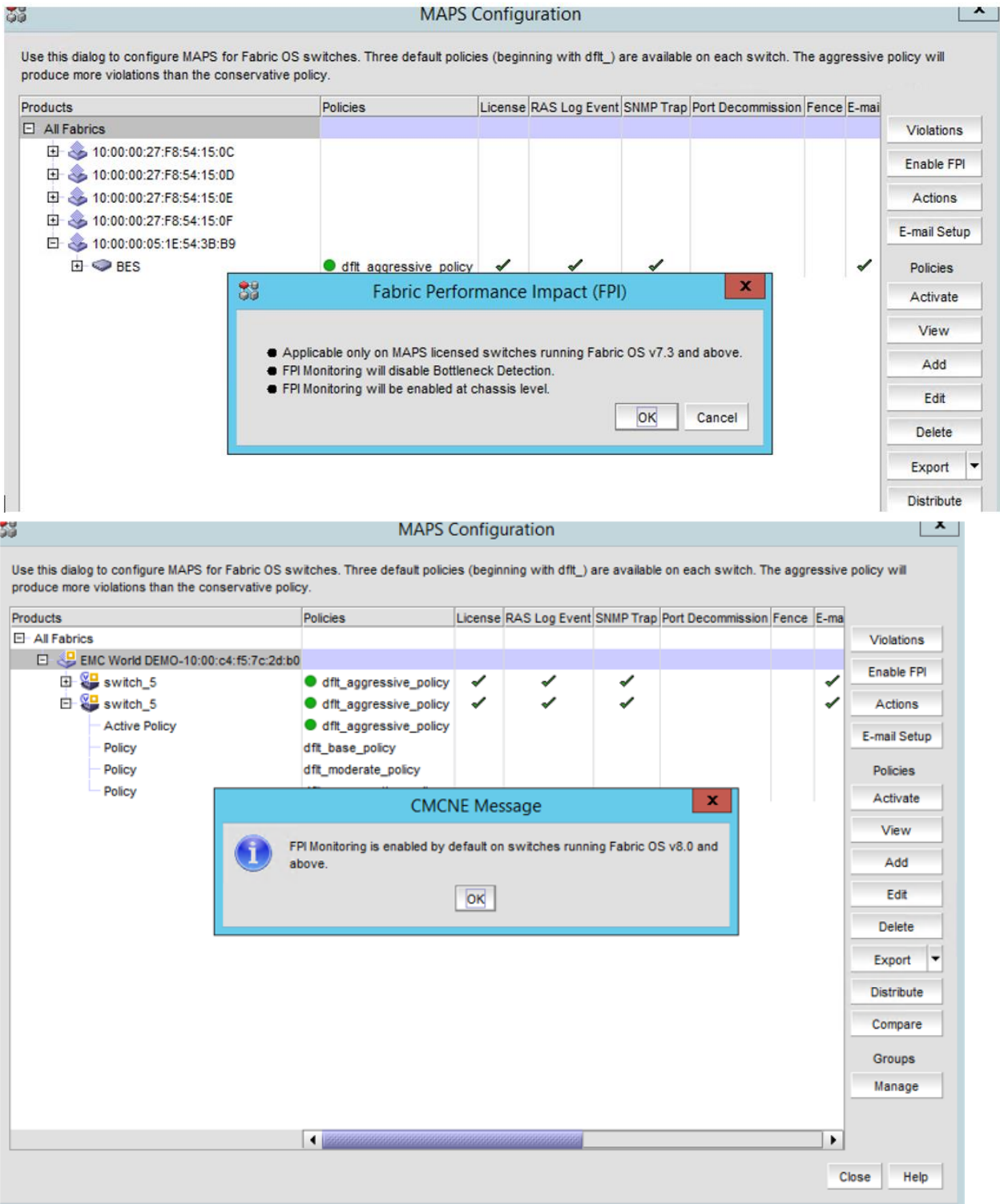
- 启用 MAPS 和 FPI

1. 单击 Monitor > Fabric Vision > MAPS > Configure



2. 突出显示结构并启用 FPI。

注意：运行 FOS 8.0 及更高版本的交换机上默认已启用 FPI。

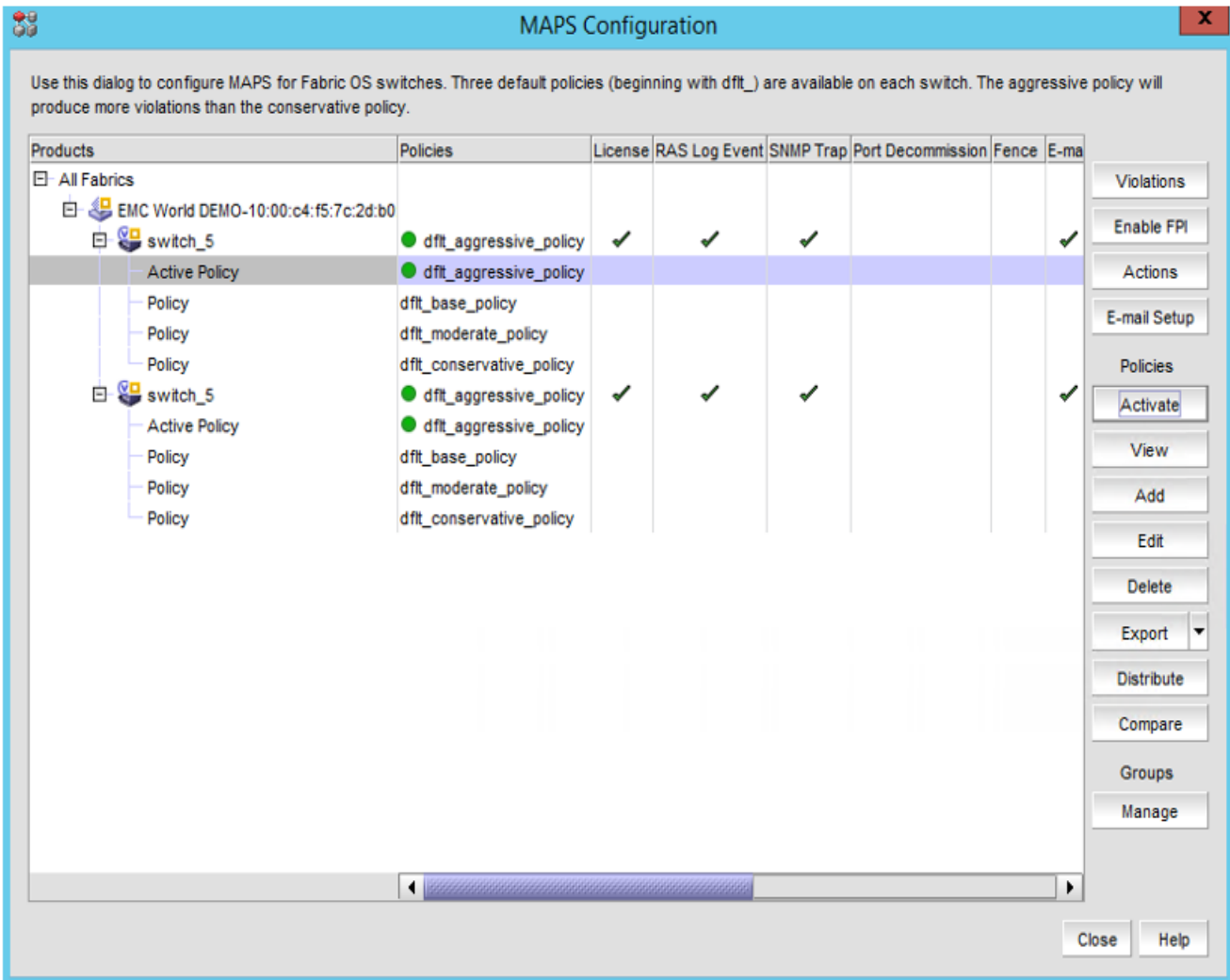


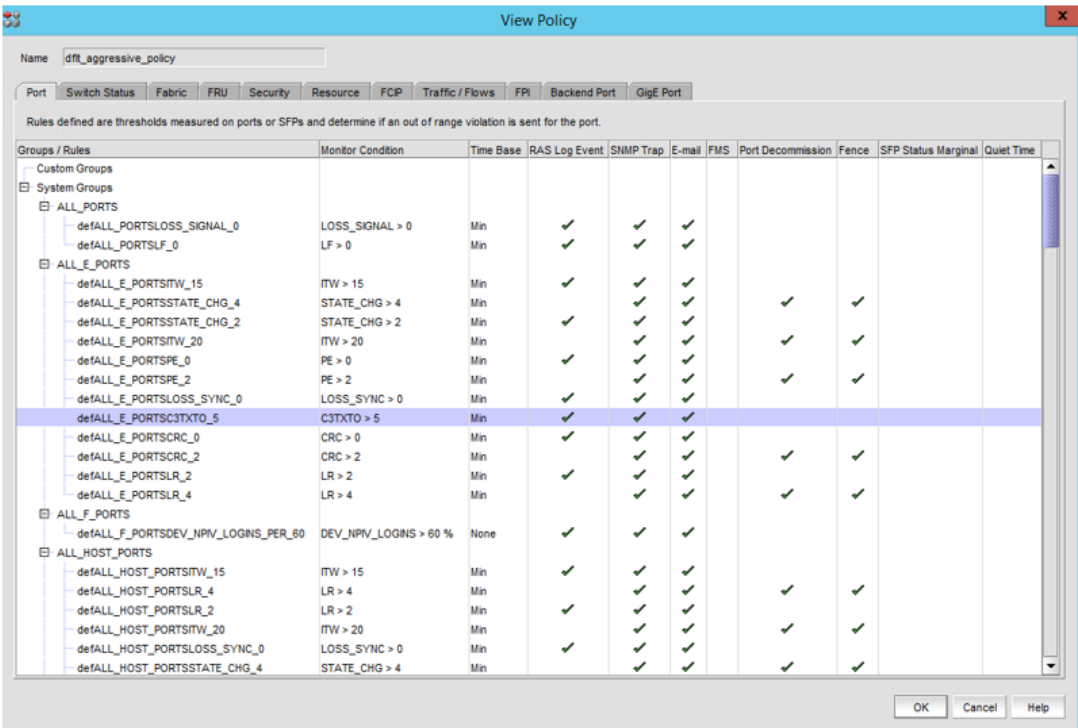
3. 在此菜单中，您可以配置结构中的每个交换机并设置所需的 MAPS 策略。

注意： CMCNE 提供预定义的策略，您可以复制它们，然后对其进行编辑。您无法编辑默认策略。有关每个策略和设置的详细信息，请参阅 MAPS 管理员指南。

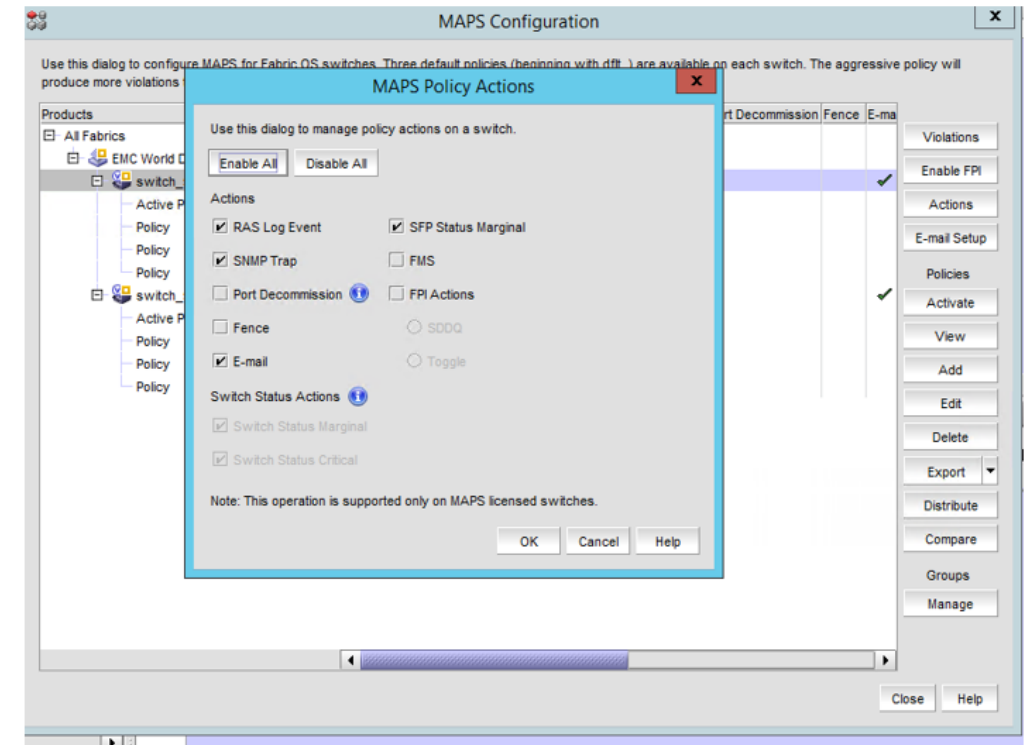
在这种情况下，我们将激活默认的积极策略。为此，请突出显示 “dflt_aggressive_policy”，然后单击 **activate**。此步骤必须在结构中要启用该策略的所有交换机上重复执行，目前您无法为整个结构启用它。

我们首先要激活积极策略，以便立即了解结构中的问题。之后，您可以调整并使用其他策略（如果我们收到太多警报）。如果单击 **View**，您可以查看每个事件的阈值。

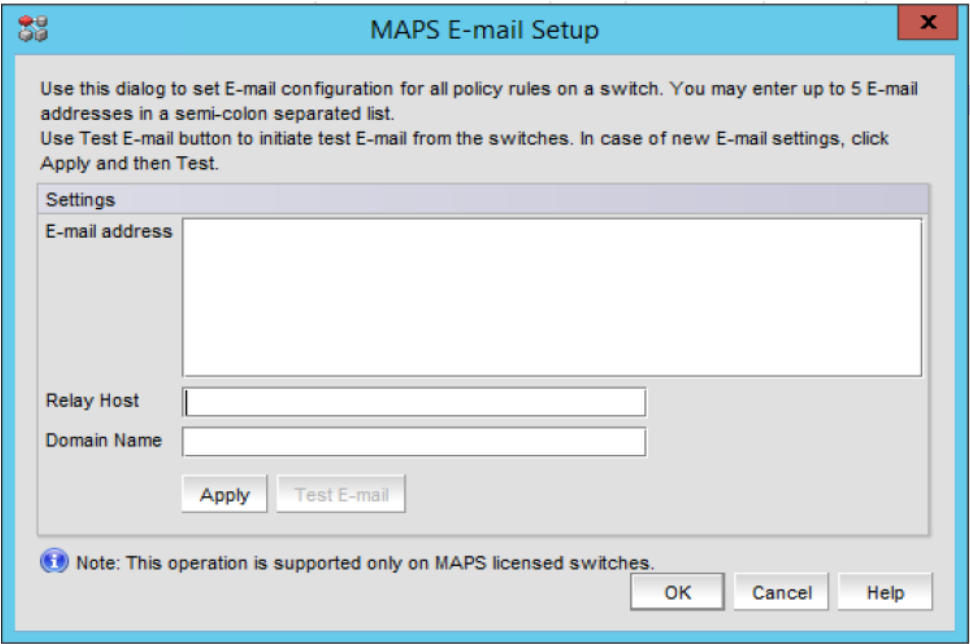




4. 突出显示某个交换机，然后单击 **Actions**。在这里，您可以决定在发生拥塞扩散时要执行的操作。对于由于超额订阅而发生拥塞扩散的特定案例分析，我们只需确保 **email** 和 **RAS log event** 已选中。



5. 如果您想要通过电子邮件接收警报，单击 **E-mail Setup** 并填写相应的字段。

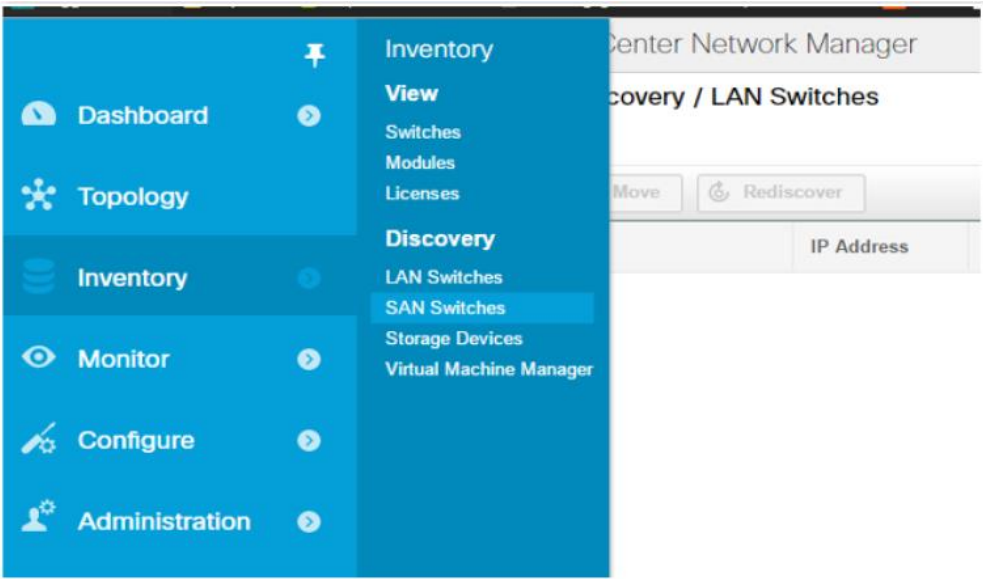


6. 确保您对结构中的**所有交换机**重复执行这些步骤。

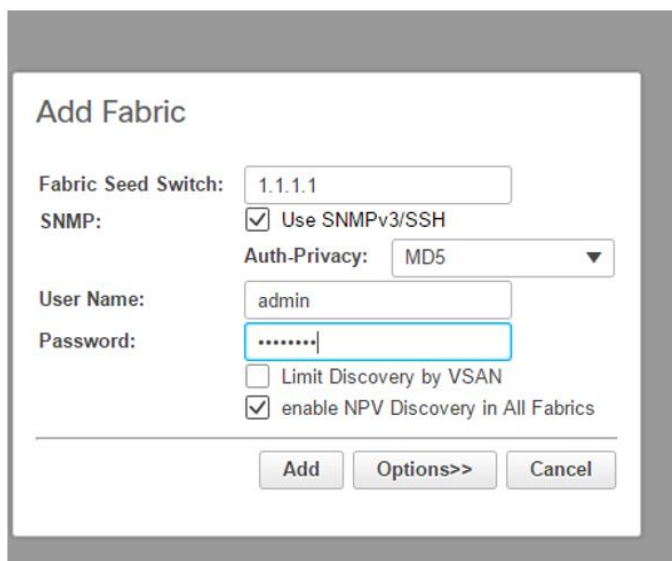
6.1.2 Cisco

- 发现结构

1. 登录到 DCNM，然后单击 Inventory > Discovery > SAN > Switches。



2. 在新窗口中，单击**加号 (+)**，填写结构中的其中一个交换机所需的信息。



The image shows a 'Add Fabric' configuration window. It contains several input fields and checkboxes. The 'Fabric Seed Switch' field is set to '1.1.1.1'. The 'SNMP' section has a checked box for 'Use SNMPv3/SSH' and a dropdown for 'Auth-Privacy' set to 'MD5'. The 'User Name' field is set to 'admin'. The 'Password' field is masked with dots. There are two checkboxes at the bottom: 'Limit Discovery by VSAN' (unchecked) and 'enable NPV Discovery in All Fabrics' (checked). At the bottom right are three buttons: 'Add', 'Options>>', and 'Cancel'.

Add Fabric

Fabric Seed Switch: 1.1.1.1

SNMP: ☒ Use SNMPv3/SSH
Auth-Privacy: MD5 ▼

User Name: admin

Password:

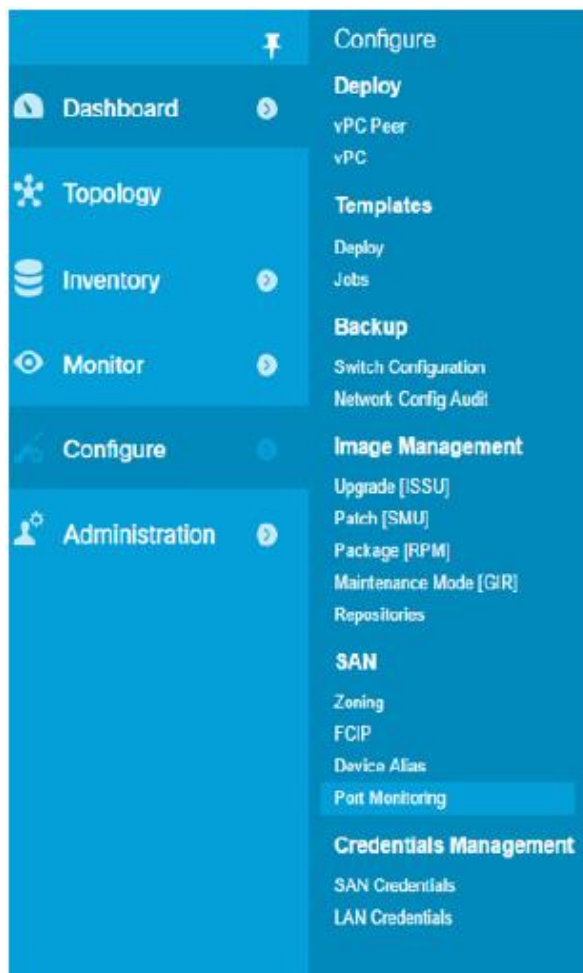
☐ Limit Discovery by VSAN
☒ enable NPV Discovery in All Fabrics

Add Options>> Cancel

3. 对所有其他结构重复此部分的操作。

启用 Cisco 端口监视 (PMON)

1. 单击 **Configure > SAN > Port Monitoring**。



2. 选择默认配置文件，然后单击 **Push to switches**。

Dashboard

Topology

Inventory

Monitor

Configure

Administration

Data Center Network Manager

Configure / SAN / Port Monitoring

default Save Save As Delete Push to switches

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. 选择所有结构，然后单击 **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.

Push

Cancel

注意：IP 地址已故意删除。

Push to switches Result

Policy: default
Port Type: All

Total 2    

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

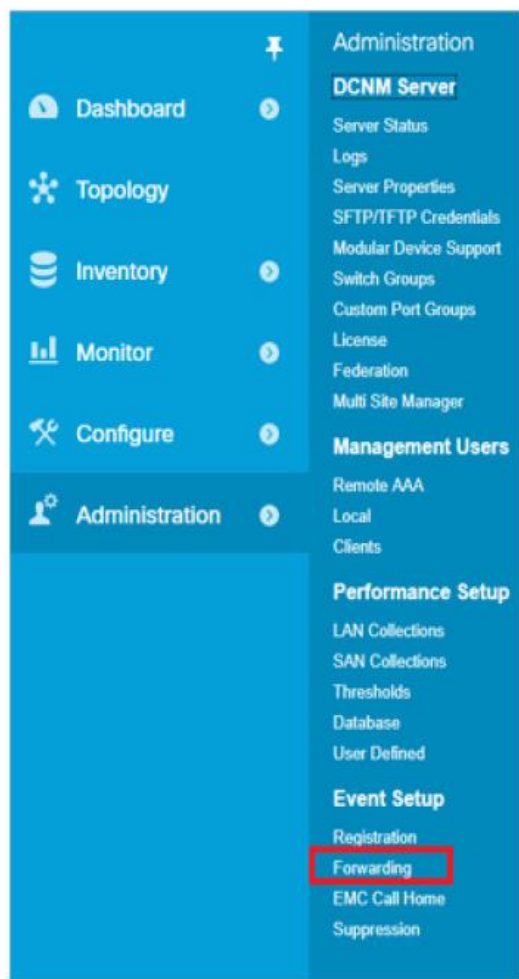
Done

Log...

Ok

Cancel

4. 通过 Cisco MDS，您可以通过 SNMP 或 Syslog 接收警报。请参阅以下配置指南，了解关于以下两项配置的信息：
<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. 要配置电子邮件主页（可选），请单击 **Administration > Event Setup**。



6. 单击加号 (+)，提供收件人电子邮件地址，然后单击 **Add**。

The image shows a screenshot of the 'Add Event Forwarder Rule' dialog box in the DCNM interface. The dialog has several fields and options: 'Forwarding Method' with radio buttons for 'E-Mail' (selected) and 'Trap'; 'Email Address' with a text box containing 'dell_emc@dell.com'; 'Forwarding Scope' with radio buttons for 'Fabric/LAN' (selected) and 'Port Groups'; 'Scope' with a dropdown menu showing 'All Fabrics'; 'VSAN Scope' with radio buttons for 'All' (selected) and 'List'; 'Source' with radio buttons for 'DCNM' (selected) and 'Syslog'; 'Type' with a dropdown menu showing 'All'; an unchecked checkbox for 'Storage Port Only'; and 'Minimum Severity' with a dropdown menu showing 'Emergency'. At the bottom right are 'Add' and 'Cancel' buttons. The 'Add' button is highlighted with a light blue background.

7. 填写 SMTP 服务器信息和发件人电子邮件地址，然后单击 **apply and test** 以确认您已收到电子邮件。

Administration / Event Setup / Forwarding

☒ Enable

SMTP Server: 1.1.1.1

From: dcnm@del.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. 确保性能监视正在运行。单击 **Administration > Server Status**。确保 **Performance Collector** 正在运行。如果没有运行，请单击 **play** 按钮来启动它。

Administration

DCNM Server

Server Status

Logs

Server Properties

SFTP/FTP Credentials

Modular Device Support

Switch Groups

Custom Port Groups

License

Federation

Multi Site Manager

Management Users

Remote AAA

Local Clients

Performance Setup

LAN Collections

SAN Collections

Thresholds

Database

User Defined

Event Setup

Registration

Forwarding

EMC Call Home

Suppression

Dashboard

Topology

Inventory

Monitor

Configure

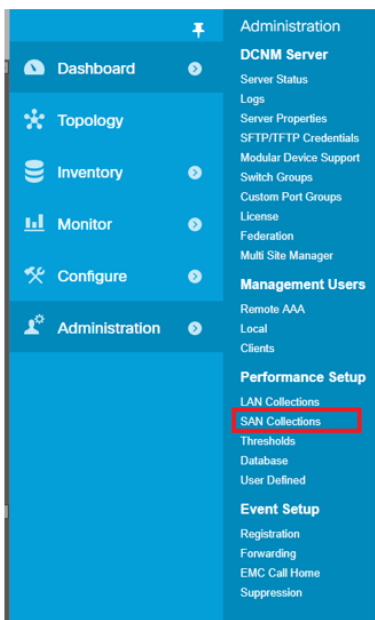
Administration

Administration / DCNM Server / Server Status

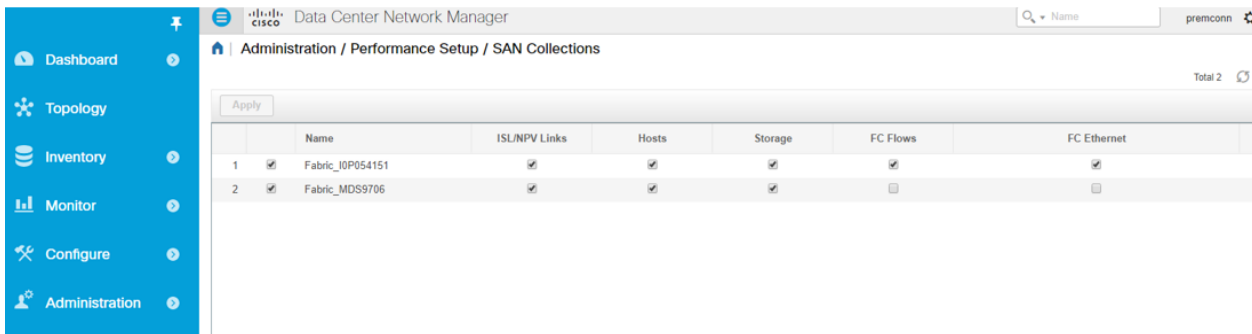
Status

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. 单击 **Administration > Performance Setup > SAN > Collections**。



10. 确保已选中想要为其收集性能统计信息的结构。Performance Collector 服务将重新启动。



参考资料

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

Cisco 慢漏设备白皮书：
<http://www.cisco.com/c/dam/en/us/products/collateral/storage-networking/mds-9700-series-multilayer-directors/whitepaper-c11-737315.pdf>

有关 VMAX 主机 I/O 限制功能的一般参考信息：
<https://community.emc.com/thread/188068?start=0&tstart=0>

Ezfio I/O 工具
<https://github.com/earlephilhower/ezfio>

拥塞扩散严重性

尽管拥塞扩散指标都非常重要，但如下一节中所述，事件发生速率会显著改变每个事件对环境的影响。让事情更复杂的是，Brocade 和 Cisco 均使用不同的严重性分类方案。因此，我们将使用以下 Dell EMC 特定分类方案并将其映射到各类交换机，如下所示：

6.1.3 Dell EMC

- 类型 1:
 - 拥塞比率大于或等于 0.2
 - 无帧丢失（丢弃）或链路重置
- 类型 2:
 - 拥塞比率大于或等于 0.2
 - 帧丢失（丢弃），但无链路重置
- 类型 3:
 - 拥塞比率大于或等于 0.2
 - 帧丢失（丢弃）和链路重置

6.1.4 Brocade

- 轻微
 - 轻微信用延迟
 - 轻微队列延迟（小于 10ms）
 - 无帧丢失（丢弃）或链路重置
- 适度
 - 中等信用延迟
 - 中等队列延迟（10ms – 80ms）
 - 帧丢失（丢弃），但无链路重置
- 严重
 - 严重信用延迟
 - 严重队列延迟（大于 80ms）
 - 帧丢失（丢弃）和部分链路重置

6.1.5 Cisco

- 级别 1: 延迟
 - 剩余信用数减少或信用不可用性持续时间短
 - 无丢弃、重新传输或链路重置
- 级别 2: 重新传输
 - 信用不可用性持续时间更长
 - 由于拥塞下降超时导致帧被丢弃, 或无信用下降超时*导致重新传输。
- 级别 3: 极大延迟
 - 信用不可用性持续时间较长 (F 端口为 1 秒, E 端口为 1.5 秒)
 - 链路重置或端口漂移

拥塞扩散术语交叉引用

指标和严重性可以合并，用于帮助确定不同类型的拥塞扩散事件。与上一节一样，Brocade 和 Cisco 都有一个单独的部分，但由于 Brocade 和 Cisco 使用术语“超额订阅”，因此本部分将首先概述该术语。

6.1.6 超额订阅

超额订阅只是“系统上的潜在需求超过系统的容量，无法满足该需求”的一种情况。其中很多人熟悉的一个示例是公路系统。如果每个人同时突然决定驾驶汽车（例如在飓风疏散事件期间），则交通将处于停滞状态。

就 FC SAN 来说，从带宽 (BW) 比率方面考虑超额订阅会很有帮助。例如，如图 3 所示，主机 1 (4 Gbps) 与存储 1 (16 GBPS) 之间的带宽比率为 1:4。因此，我们可以说主机 1 是 4:1 超额订阅。将它与主机 2 (16 Gbps) 和存储 2 (16 Gbps) 之间的带宽比率（即 1:1）进行比较。考虑主机和存储都使用 32 Gbps ISL 进行访问的情况，您可以看到，主机 2 和存储 2 之间没有超额订阅。在这种情况下，我们说主机 2 和存储 2 为非超额订阅。

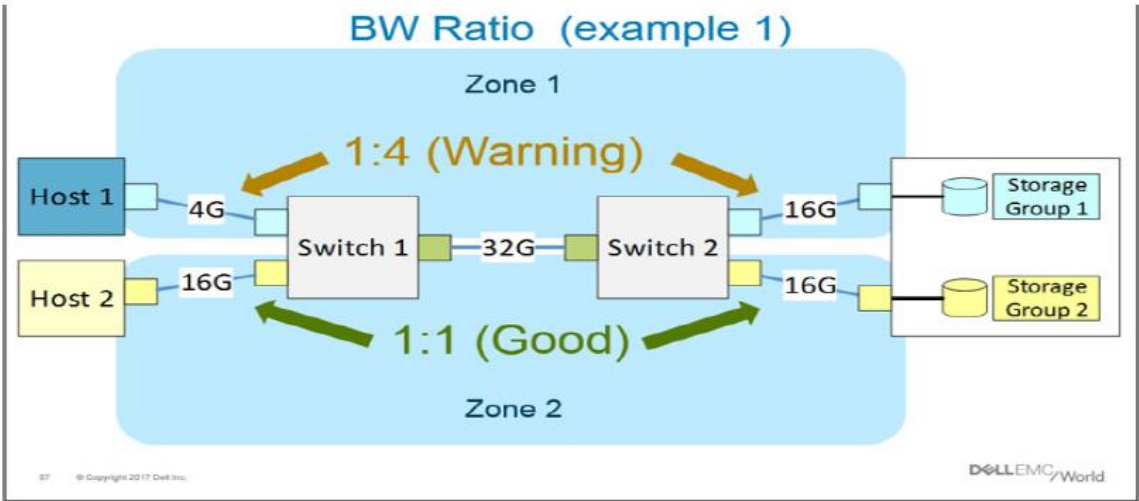


图 19 带宽比率 — 示例 1

需要注意的是，在计算超额订阅时（如图 19 所示），带宽比率是通过添加要考虑的接口的带宽来计算。乍看起来，您可能认为我们是使用 16 Gbps HBA 访问 8 Gbps 存储，但是由于实际上有三个存储接口，因此我们是使用一个 16 Gbps HBA 访问 24 Gbps 存储。因此，主机是 3:2 超额订阅。

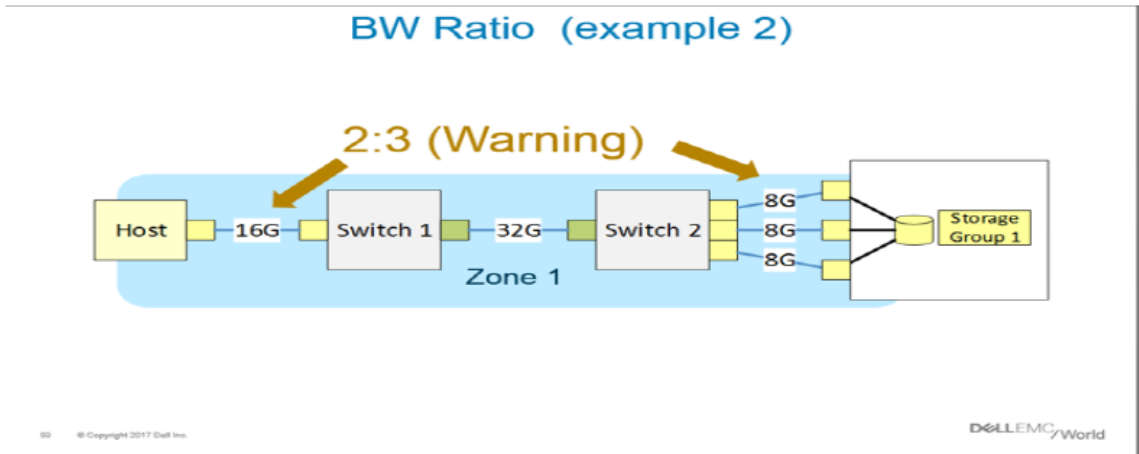


图20 带宽比率 — 示例2

在前面的 2 个示例中，ISL 的带宽始终大于或等于终端设备可以支持的带宽量。通常情况并非如此。如图 20 所示，主机实际上是 3:4 订阅不足，但由于 ISL 仅为 16 Gbps，所以终端设备与将使用的 ISL 之间存在超额订阅，您可以说 ISL 是 3:2 超额订阅。

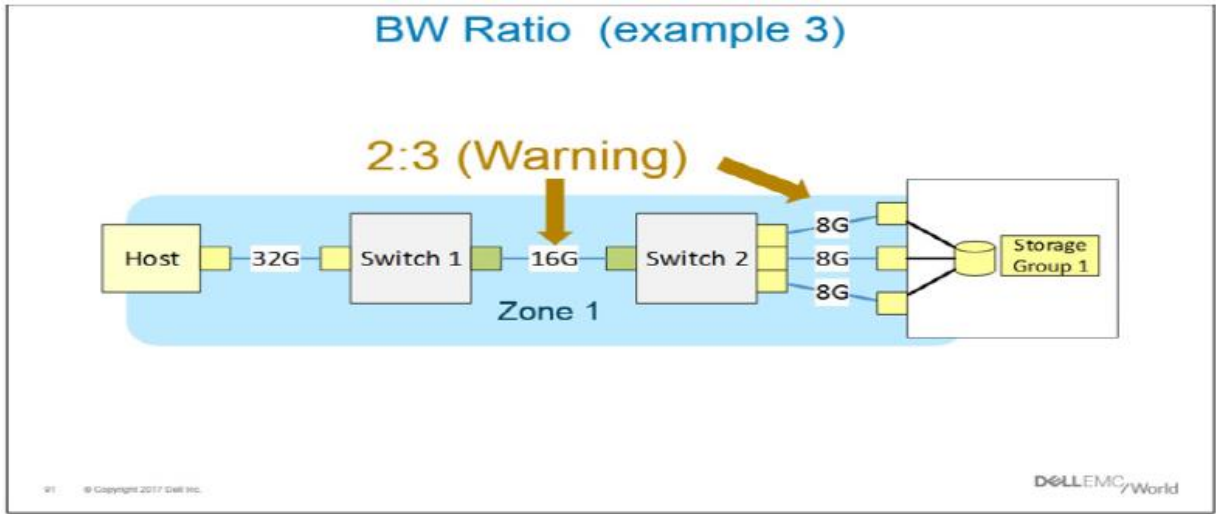


图21 带宽比率 — 示例3

6.1.7 Brocade
Brocade 定义了三个不同的拥塞扩散事件类别：

- 超额订阅

如上一节中所定义（如上）。

• 行为异常的设备

未以足够快的速度释放信用以维持线速的终端设备或 ISL。例如，如果终端设备协商了 16 Gbps 的链路速度，但无法以允许其接收 16 Gbps 数据的速率恢复信用，那么您可能有一台行为异常的设备。这些类型的设备也称为“慢漏”。需要指出的是，设备可能因许多原因（包括驱动程序问题，或者在 ISL 中，因为端口正在经历拥塞扩散效应）而出现行为异常。

• 丢失信用

丢失信用场景意味着，出于某种原因（例如，通常偶尔出现位错误），给定链路上的一台或两台设备认为它们拥有较少的传输信用，而实际上并非如此。出现这种情况的一个原因是，因为位错误而损坏 R_RDY。如果这种情况时常发生，性能就会随着时间的推移开始下降，并慢慢降低 SAN 传输数据的能力。此问题将在知识库文章 464245（位错误及其影响）中更详细地进行探讨。

Brocade 拥塞扩散术语交叉引用

对于 Brocade，将所有信息结合在一起得出以下特定于 Brocade 的拥塞扩散术语交叉引用。

原因	轻微	中等	严重
超额订阅 ¹	1. 设备端口具有高带宽 2. ISL 端口的短暂信用延迟 3. 无帧丢失或链路重置	1. 设备端口具有高带宽 2. ISL 端口的中等信用延迟 3. ISL 端口的队列延迟在 10 毫秒到 80 毫秒之间 4. 无帧丢失或链路重置	1. 设备端口具有高带宽 2. ISL 端口的长时间信用延迟 3. ISL 端口的队列延迟超过 80 毫秒 4. 上游 (ISL) 端口的帧丢失（说明队列延迟为 220 毫秒到 500 毫秒） 5. 无链路重置。
行为异常的设备	1. 设备端口和上游 ISL 端口的短暂信用延迟 2. 设备端口和上游 ISL 端口的队列延迟短于 10 毫秒 3. 无帧丢失或链路重置	1. ISL 端口和上游 ISL 端口的中等信用延迟 2. 设备端口和上游 ISL 端口的队列延迟在 10 毫秒到 80 毫秒之间 3. 无帧丢失或链路重置	1. 设备端口和上游 ISL 端口的长时间信用延迟 2. 设备端口和上游 ISL 端口的队列延迟超过 80 毫秒 3. 设备或上游 (ISL) 端口的帧丢失（说明队列延迟为 220 毫秒到 500 毫秒） 4. ISL 端口链路重置（说明信用停滞超过 2 秒）

丢失信用 ²	1. 端口的短暂信用延迟 2. 端口或端口上游的队列延迟短于 10 毫秒 3. 无帧丢失或链路重置	1. 端口的中等信用延迟 2. 端口或端口上游的队列延迟在 10 毫秒到 80 毫秒之间 3. 无帧丢失或链路重置	1. 端口的长时间信用延迟 2. 端口或端口上游的队列延迟超过 80 毫秒 3. 端口或端口上游的帧丢失（说明信用停滞为 220 毫秒到 500 毫秒） 4. 端口或端口上游的链路重置（说明信用停滞超过 2 秒）
¹ 由于超额订阅造成严重拥塞的情况很少或极少发生。 ² 丢失信用的原因通常为传输错误，如 ITW、CRC 或其他与信号相关的问题。			

6.1.8 Cisco

Cisco 定义了两个不同的拥塞扩散事件类别：

- 超额订阅
- 如上面所定义。
- 信用不足

未以足够快的速度释放信用以维持线速的终端设备或 ISL。例如，如果终端设备协商了 16 Gbps 的链路速度，但无法以允许其接收 16 Gbps 数据的速率恢复信用，那么您可能有一台行为异常的设备。这些类型的设备也称为“慢漏”。需要指出的是，设备可能因许多原因（包括驱动程序问题，或者在 ISL 中，因为端口正在经历拥塞扩散效应）而出现行为异常。

- Cisco 拥塞扩散术语交叉引用

将针对 Cisco 的所有信息组合在一起得出以下特定于 Cisco 的拥塞扩散术语交叉引用。

拥塞类型	级别 1：延迟	级别 2：重新传输	级别 3：极大延迟
超额订阅	1. 终端设备端口上的链路利用率高 2. 终端设备端口无 B2B 信用不足 3. ISL 的拥塞扩散 4. 无帧丢失或链路重置	由于超额订阅造成重新传输或极大延迟的情况很少或极少发生	
信用不足	1. 终端设备端口上的链路利用率低	1. 终端设备端口上的链路利用率低	1. 不会将任何帧传输到终端设备

	2. 剩余信用数减少或信用不可用性持续时间短 3. ISL 的拥塞扩散 4. 无丢弃、重新传输或链路重置	2. 信用不可用性持续时间更长 3. ISL 的拥塞扩散 4. 由于拥塞下降超时或无信用下降超时*导致重新传输，以致帧被丢弃（但不会重置链路）	2. 信用不可用性持续时间较长（F 端口为 1 秒，E 端口为 1.5 秒） 3. ISL 的严重拥塞 4. 链路重置或端口漂移
*默认配置：拥塞下降超时 — 500 毫秒，无信用下降超时 — 关闭 可配置选项：拥塞下降超时 — 100-500 毫秒，无信用下降超时 — 1-500 毫秒 推荐配置：拥塞下降超时 — 200 毫秒，无信用下降超时 — 50 毫秒			