用户指南

聚合网络适配器

41*xxx* Series

第三方信息由 Dell 提供给您。

AH0054602-05 M

2019 年 10 月 16 日

Marvell. 快步向前

有关更多信息，请访问网址：http://www.marvell.com

目录

15 Windows Server 2016

16 Windows Server 2019

图片列表

表格列表

前言

本前言列出了支持的产品，指定了读者对象，说明了本指南中使用的排版惯例，并介绍了法律声明。

支持的产品

注

QConvergeConsole® (QCC) GUI 是*唯一*涵盖所有 Marvell® FastLinQ® 适配器的 GUI 管理工具。 QLogic Control Suite™ (QCS) GUI 不再支持 FastLinQ 45000 系列适配器以及基于 57*xx*/57*xxx* 控制器的适配器，已经被 QCC GUI 管理工具取代。 QCC GUI 为所有 Marvell 适配器提供单面板 GUI 管理。

在 Windows® 环境中，当您运行 QCS CLI 和管理代理安装程序时，将从系统中卸载 QCS GUI （如果系统上已安装）及任何相关组件。要获取新 GUI，请从 Marvell 网站为适配器下载 QCC GUI （请参阅第 xx 页上 “ 下载更新和文档 ”）。

本用户指南介绍以下 Marvell 产品：

- QL41112HFCU-DE 10Gb 聚合网络适配器，全高支架
- QL41112HLCU-DE 10Gb 聚合网络适配器，薄型支架
- QL41132HFRJ-DE 10Gb NIC 适配器，全高支架
- QL41132HLRJ-DE 10Gb NIC 适配器，薄型支架
- QL41132HQCU-DE 10Gb NIC 适配器
- QL41132HQRJ-DE 10Gb NIC 适配器
- QL41154HQRJ-DE 10Gb 聚合网络适配器
- QL41154HQCU-DE 10Gb 聚合网络适配器
- QL41162HFRJ-DE 10Gb 聚合网络适配器，全高支架
- QL41162HLRJ-DE 10Gb 聚合网络适配器，薄型支架
- QL41162HMRJ-DE 10Gb 聚合网络适配器
- QL41164HMCU-DE 10Gb 聚合网络适配器
- QL41164HMRJ-DE 10Gb 聚合网络适配器

- QL41164HFRJ-DE 10Gb 聚合网络适配器，全高支架
- QL41164HFRJ-DE 10Gb 聚合网络适配器，薄型支架
- QL41164HFCU-DE 10Gb 聚合网络适配器，全高支架
- QL41232HFCU-DE 10/25Gb NIC 适配器，全高支架
- QL41232HLCU-DE 10/25Gb NIC 适配器，薄型支架
- QL41232HMKR-DE 10/25Gb NIC 适配器
- QL41232HQCU-DE 10/25Gb NIC 适配器
- QL41262HFCU-DE 10/25Gb 聚合网络适配器，全高支架
- QL41262HLCU-DE 10/25Gb 聚合网络适配器，薄型支架
- QL41262HMCU-DE 10/25Gb 聚合网络
- QL41262HMKR-DE 10/25Gb 聚合网络适配器
- QL41264HMCU-DE 10/25Gb 聚合网络适配器

读者对象

本指南面向负责对安装在 Dell® PowerEdge® 服务器 （在 Windows®、 Linux® 或 VMware® 环境中）上的适配器进行配置和管理的系统管理员和其他技术人员。

本指南的内容

前言之后，本指南的其余部分划分为以下章节和附录：

- 第 1 章 产品概览提供产品的功能说明、功能列表以及适配器规格。
- 第 2 章 硬件安装介绍如何安装适配器，包括系统要求列表和预安装核查表。
- 第 3 章 驱动程序安装介绍在 Windows、 Linux 和 VMware 中安装适配器驱动程序。
- 第 4 章 升级固件介绍如何使用 Dell Update Package (DUP) 升级适配器固件。
- 第 5 章 适配器预引导配置介绍如何使用人机界面基础设施 (HII) 应用程序执行预引导适配器配置任务。
- 第 6 章 从 SAN 引导配置涵盖用于 iSCSI 和 FCoE 的从 SAN 引导配置。
- 第 7 章 RoCE 配置介绍如何配置适配器、以太网交换机和主机以使用基于聚合以太网的 RDMA (RoCE)。
- 第 8 章 iWARP 配置 提供在 Windows、 Linux 和 VMware ESXi 6.7 系统中配置互联网广域 RDMA 协议 (iWARP) 的步骤。
- 第 9 章 iSER 配置介绍如何为 Linux RHEL、SLES、Ubuntu 和 ESXi 6.7 配置 RDMA 的 iSCSI 扩展 (iSER)。
- 第 10 章 iSCSI 配置介绍适用于 Windows 和 Linux 的 iSCSI 引导和 iSCSI 卸载。
- 第 11 章 FCoE 配置涵盖配置 Linux FCoE 卸载。

- 第 12 章 SR-IOV 配置提供在 Windows、Linux 和 VMware 系统中配置单根输入 / 输出虚拟化 (SR-IOV) 的步骤。
- 第 13 章 使用 RDMA 的 NVMe-oF 配置演示如何在简单的网络上为 41*xxx* Series Adapters 配置 NVMe-oF。
- 第 14 章 VXLAN 配置介绍如何为 Linux、 VMware 和 Windows Server 2016 配置 VXLAN。
- 第 15 章 Windows Server 2016 介绍 Windows Server 2016 功能。
- 第 16 章 Windows Server 2019 介绍 Windows Server 2019 功能。
- 第 17 章 故障排除介绍各种故障排除方法和资源。
- 附录 A 适配器 LED 列出适配器 LED 及其意义。
- 附录 B 电缆和光学模块列出41*xxx* Series Adapters支持的电缆、光学模块和交换机。
- 附录 C Dell Z9100 交换机配置介绍如何配置 Dell Z9100 交换机端口以建立 25Gbps 连接。
- 附录 D 功能约束提供有关当前版本中所实现功能约束的信息。
- 附录 E 修订历史介绍指南的此修订版所做的更改。

本指南的最后部分是术语表。

说明文件惯例

本指南使用以下说明文件惯例：

- **注** 提供额外的信息。
- **小心** 不带警报符号，表示存在可能导致设备损坏或数据丢失的危险。
- ⚠ **小心** 带警报符号，表示存在可能造成轻度或中度伤害的危险。
- ⚠ **警告** 表示存在可能造成严重伤害或死亡的危险。
- 蓝色字体的文字表示至本指南中的插图、表格或章节的超链接 （跳转），至网站的链接以下划线蓝色文字显示。例如：
  - 表 9-2 列出与用户界面和远程代理有关的问题。
  - 请参阅第 6 页上的 “ 安装核查表 ”。
  - 有关更多信息，请访问 www.marvell.com。

- **黑体**文字表示用户界面元素，如菜单项、按钮、复选框或列标题。例如：
  - 单击 **Start**（开始）按钮，指向 **Programs**（程序），指向 **Accessories**（附件），然后单击 **Command Prompt**（命令提示符）。
  - 在 **Notification Options**（通知选项）下，选中 **Warning Alarms**（警报）复选框。
- `Courier` 字体文本表示文件名、目录路径或命令行文字。例如：
  - 要从文件结构的任何地方返回到根目录：
    键入 **`cd/ root`** 并按 ENTER 键。
  - 发出以下命令 **`sh ./install.bin`**。
- 键盘的键名和击键用 UPPERCASE（大写字母）表示：
  - 按 CTRL+P。
  - 按向上箭头键。
- *斜体*文字表示术语、强调、变量或说明文件标题。例如：
  - 哪些是*快捷键*?
  - 要输入日期，键入 *mm/dd/yyyy*（其中 *mm* 是月， *dd* 是日， *yyyy* 是年）。
- 引号中的主题标题表示本手册内或的联机帮助的相关主题；在本说明文件中，手册或联机帮助又称为*帮助系统*。
- 命令行界面 (CLI) 命令语法惯例包括以下内容：
  - 纯文本表示必须按照所示键入的项目。例如：
    - **`qaucli -pr nic -ei`**
  - < >（尖括号）表示必须指定其值的变量。例如：
    - `<serial_number>`

    注

    仅适用于 CLI 命令，变量名称始终使用尖括号而不是*斜体*表示。
  - [ ]（方括号）表示可选的参数。例如：
    - `[<file_name>]` 表示指定文件名，或省略以选择默认文件名。

- | （竖线）表示互斥的选项；只能选择一个选项。例如：
  - on|off
  - 1|2|3|4
- ... （省略号）表示前面的项可重复。例如：
  - x... 表示 x 的*一个*或多个实例。
  - [x...] 表示 x 的*零个*或多个实例。
- ⋮ （竖省略号）命令示例输出中的垂直省略号指示重复输出数据部分被有意省略的位置。
- ( ) （圆括号）和 { } （大括号）用来避免逻辑模糊不清。例如：
  - a|b c 模糊不清
    {(a|b) c} 表示 a 或 b，后跟 c
    {a|(b c)} 表示 a 或 b c

下载更新和文档

Marvell 网站提供产品固件、软件和文档的定期更新。

要下载 Marvell 固件、软件和文档：

1. 请访问 www.marvell.com。
2. 指向 **Support** （支持），然后在 **Driver Downloads** （驱动程序下载）下单击 **Marvell QLogic/FastLinQ Drivers** （QLogic/FastLinQ 驱动程序）。
3. 在驱动程序和文档页面上，单击 **Adapters** （适配器）。
4. 单击相应的按钮以 **by Model** （按型号）或 **by Operating System** （按操作系统）搜索。
5. 要定义搜索，请单击每个选择列中的项目，然后单击 **Go** （前往）。
6. 找到所需的固件、软件或文档，然后单击该项目的名称或图标以下载或打开该项目。

法律声明

本节中涵盖的法律声明包括激光安全 （FDA 公告）、机构认证和产品安全合规性。

激光安全 - FDA 公告

本产品符合 DHHS 规则 21CFR I 章 J 节的规定。本产品的设计和生产符合 IEC60825-1 中有关激光产品安全标签的规定。

I 类激光产品

1 类 激光产品	**小心** — 在打开时存在 1 类激光辐射 请勿直视光学仪器
Appareil laser de classe 1	**Attention**—Radiation laser de classe 1 Ne pas regarder directement avec des instruments optiques
Produkt der Laser Klasse 1	**Vorsicht**—Laserstrahlung der Klasse 1 bei geöffneter Abdeckung Direktes Ansehen mit optischen Instrumenten vermeiden
Luokan 1 Laserlaite	**Varoitus**—Luokan 1 lasersäteilyä, kun laite on auki Älä katso suoraan laitteeseen käyttämällä optisia instrumenttej

机构认证

以下章节概述对 41*xxx* Series Adapters 执行的 EMC 和 EMI 检验规格，验证其是否符合辐射排放、抗辐射干扰性以及产品安全标准。

EMI 和 EMC 要求

FCC 第 15 部分符合性：A 级

FCC 符合性信息声明：本设备符合 FCC 规则第 15 部分的规定。操作应符合以下两个条件：(1) 该设备不得造成有害干扰； (2) 该设备必须能够承受接收到的任何干扰，包括可能导致不良操作的干扰。

ICES-003 符合性：A 级

本 A 级数字装置符合加拿大 ICES-003 的规定。 Cet appareil numériqué de la classe A est conformé à la norme NMB-003 du Canada.

CE 标记 2014/30/EC、 2014/35/EU EMC 指令符合性：

EN55032:2012/ CISPR 32:2015 A 级

EN55024: 2010
EN61000-3-2：谐波电流放射
EN61000-3-3：电压波动和闪动

抗扰性标准
EN61000-4-2：ESD
EN61000-4-3：RF 电磁场
EN61000-4-4：快速瞬变 / 猝变
EN61000-4-5：快速电涌常见 / 差动
EN61000-4-6：RF 传导敏感度
EN61000-4-8：电力频率磁场
EN61000-4-11：电压骤降和干扰

VCCI：2015-04; A 级

AS/NZS ； CISPR 32：2015 A 级

CNS 13438：2006 A 级

KCC：A 级

经过韩国 RRA A 级认证

产品名称 / 型号：聚合网络适配器和智能以太网适配器
证书持有方：QLogic Corporation
制造日期：请参阅产品上的日期代码
原始制造商 / 国家 （地区）：QLogic Corporation/ 美国

A 级设备
（商用信息 / 电讯设备）

由于本设备已就其商用性执行了 EMC 注册，因此要求销售方和 / 或购买方对此引起充分注意，倘若发生错误的销售或购买行为，要求立即将其更换成家用型设备。

韩国语言格式 - A 급

A급 기기 (업무용 정보통신기기)

이 기기는 업무용으로 전자파적합등록을 한 기기이오니 판매자 또는 사용자는 이 점을 주의하시기 바라며, 만약 잘못판매 또는 구입하였을 때에는 가정용으로 교환하시기 바랍니다.

VCCI：A 级

根据 Voluntary Control Council for Interference（干扰自愿控制委员会）(VCCI) 的标准，该设备是 A 级产品。如果在家庭环境中使用该设备，可能会发生无线电干扰，在这种情况下，用户可能需要采取纠正措施。

この装置は、クラスA情報技術装置です。この装置を家庭環境で使用すると電波妨害を引き起こすことがあります。この場合には使用者が適切な対策を講ずるよう要求されることがあります。 ＶＣＣＩ－Ａ

产品安全符合性

UL、 cUL 产品安全：

UL 60950-1 （第二版） A1 + A2 2014-10-14
CSA C22.2 No.60950-1-07 （第二版） A1 +A2 2014-10

只能用于所列 ITE 或等价对象。

符合 21 CFR 1040.10 和 1040.11、 2014/30/EU、 2014/35/EU。

2006/95/EC 低电压规程：

TUV EN60950-1:2006+A11+A1+A12+A2 第二版
TUV IEC 60950-1: 2005 第二版 Am1: 2009 + Am2: 2013 CB

根据 IEC 60950-1 第二版经过 CB 认证

1 产品概览

本章提供 41*xxx* 系列适配器的以下信息：

功能说明

Marvell FastLinQ 41000 系列适配器包括 10 和 25Gb 聚合网络适配器以及智能以太网适配器，旨在为服务器系统执行加速的数据网络。 41000 系列适配器包括一个 10/25Gb 以太网 MAC （具有全双工能力）。

利用操作系统的组合功能，可将网络分割成虚拟 LAN (vLANs)，以及将多个网络适配器组合到各个组中，以便提供网络负载平衡和容错。有关组合的更多信息，请参阅操作系统说明文件。

功能

41*xxx* 系列适配器提供以下功能。并非所有适配器均具备所有功能：

- NIC 分区 (NPAR)
- 单芯片解决方案：
  - 10/25Gb MAC
  - 用于直连式铜缆 (DAC) 收发器连接的 SerDes 接口
  - PCI Express® (PCIe®) 3.0 x8
  - 具备零拷贝功能的硬件
- 性能特点：
  - TCP、 IP、 UDP 校验和卸载
  - TCP 分段卸载 (TSO)
  - 大段卸载 (LSO)
  - 普通分段卸载 (GSO)

- ❑ 大量接收卸载 (LRO)
- ❑ 接收分段结合 (RSC)
- ❑ Microsoft® 动态虚拟机队列 (VMQ) 和 Linux 多队列

- 自适应中断：
  - ❑ 发送方 / 接收方缩放 (TSS/RSS)
  - ❑ 使用通用路由封装 (NVGRE) 和虚拟 LAN (VXLAN) L2/L3 GRE 隧道流量进行网络虚拟化无状态卸载[1]
- 可管理性：
  - ❑ 系统管理总线 (SMB) 控制器
  - ❑ 符合*高级配置和电源接口* (ACPI) 1.1a 标准 （多电源模式）
  - ❑ 网络控制器边带接口 (NC-SI) 支持
- 高级网络特性：
  - ❑ 巨型帧 （多达 9,600 个字节）。 OS 和链路伙伴必须支持巨型帧。
  - ❑ 虚拟 LAN (VLAN)
  - ❑ 流控制 (IEEE Std 802.3x)
- 逻辑链路控制 （IEEE 标准 802.2）
- 高速芯片搭载精简指令集计算机 (RISC) 处理器
- 集成式 96KB 帧缓冲区存储区 （并非适用于所有型号）
- 1,024 分类过滤器 （并非适用于所有型号）
- 通过 128 位散列硬件功能支持多播地址
- 对 VMDirectPath I/O 的支持

  FastLinQ 41*xxx* 系列适配器 支持在 Linux 和 ESX 环境中的 VMDirectPath I/O。 Windows 环境下不支持 VMDirectPath I/O。

  FastLinQ 41*xxx* 系列适配器 可以分配给虚拟机进行 PCI pass-through 操作。但是，由于功能级别的依赖性，与适配器相关联的所有 PCIe 功能都必须分配给同一个虚拟机。不支持跨虚拟机监控程序和 / 或一个或多个虚拟机共享 PCIe 功能。
- 串行 NVRAM 闪存
- *PCI 电源管理接口* (v1.1)
- 64 位基本地址寄存器 (BAR) 支持

[1] 此功能要求 OS 或虚拟机监控程序支持才能使用卸载。

- EM64T 处理器支持
- iSCSI 和 FCoE 引导支持[2]

适配器规格

41*xxx* 系列适配器规格包含适配器的物理特性和标准符合性参考。

物理特性

41*xxx* 系列适配器 是标准的 PCIe 卡，并附带一个全高或薄型支架以在标准 PCIe 插槽中使用。

标准规格

支持的标准规格包括：

- *PCI Express 基本规格*，修订版 3.1
- *PCI Express 卡机电规格*，修订版 3.0
- *PCI 总线电源管理接口规格*，修订版 1.2
- IEEE 规格：
  - *802.1ad* (QinQ)
  - *802.1AX*（链路聚合）
  - *802.1p*（优先级编码）
  - *802.1q* (VLAN)
  - *以太网的 802.3-2015 IEEE 标准*（流控制）
  - *802.3-2015 Clause 78 节能以太网 (EEE)*
  - *1588-2002 PTPv1*（精确时间协议）
  - *1588-2008 PTPv2*
- IPv4 (RFQ 791)
- IPv6 (RFQ 2460)

[2] SR-IOV VF 的硬件支持限制有所不同。该限制在某些 OS 环境中可能较低；请参阅您的 OS 的相应章节。

2 硬件安装

本章提供以下硬件安装信息：

系统要求

安装 Marvell 41*xxx* 系列适配器 前，请确认您的系统满足表 2-1 和表 2-2 所示的硬件和操作系统要求。有关支持的操作系统的完整列表，请访问 Marvell 网站。

表 2-1. 主机硬件要求

硬件	要求
体系结构	可满足操作系统要求的 IA-32 或 EMT64
PCIe	PCIe Gen 2 x8 (2x10G NIC) PCIe Gen 3 x8 (2x25G NIC) PCIe Gen 3 x8 或更快的插槽支持全双端口 25Gb 带宽。
内存	8GB RAM （最少）
电缆和光学模块	41*xxx* 系列适配器 已通过与各种 1G、 10G 和 25G 电缆和光学模块的互操作性测试。请参阅 第 289 页上 “ 测试的电缆和光学模块 ”。

表 2-2. 最低主机操作系统要求

操作系统	要求
Windows Server	2012 R2、 2019
Linux	RHEL® 7.6、 7.7、 8.0、 8.1 SLES® 12 SP4、 SLES 15、 SLES 15 SP1 CentOS 7.6
VMware	vSphere® ESXi 6.5 U3 和 vSphere ESXi 6.7 U3
XenServer	Citrix 虚拟机监控程序 8.0 7.0、 7.1

注

表 2-2 说明最低主机 OS 要求。有关支持的操作系统的完整列表，请访问 Marvell 网站。

安全预防措施

⚠ 警告

安装适配器的系统的操作电压可能会有致命危险。打开系统外壳之前，请遵从以下预防措施以保护您自己并避免损坏系统组件。

- 除去手上和手腕上的任何金属物体或首饰。
- 确保仅使用绝缘工具或非导电工具。
- 在触摸内部组件之前，确认系统电源已关闭并已拔掉电源线。
- 在不受静电干扰的环境中安装或卸下适配器。使用正确接地的腕带或其他人体防静电设备，强烈建议使用防静电地垫。

预安装核查表

安装适配器之前，请完成以下操作：

1. 确认系统满足 第 4 页上 “ 系统要求 ” 中列出的硬件和软件要求。
2. 确认系统使用最新的 BIOS。

注

如果您从 Marvell 网站获取适配器软件，请确认适配器驱动程序文件的路径。

3. 如果系统正在运行，请将其关闭。
4. 系统关闭后，断开电源并拔下电源线。
5. 将适配器从其运输包装中取出并放在防静电表面上。
6. 检查适配器，特别是边缘连接器上是否有明显的损坏痕迹。切勿尝试安装损坏的适配器。

安装适配器

以下说明适用于在大多数系统中安装 Marvell 41*xxx* 系列适配器。有关执行这些任务的详细信息，请参阅随系统提供的手册。

要安装适配器：

1. 复查 第 5 页上 “ 安全预防措施 ” 和 第 6 页上 “ 预安装核查表 ”。安装适配器前，确保系统电源已关闭而且电源线已从电源插座上拔下，并且遵守适当的电接地步骤。
2. 打开系统机箱，然后选择符合适配器大小的插槽，可以是 PCIe Gen 2 x8 或 PCIe Gen 3 x8。较窄的适配器可插入更宽的插槽中 （x8 的适配器可插入 x16 的插槽中），但较宽的适配器不能插入更窄的插槽中 （x8 的适配器不能插入 x4 的插槽中）。如果不知道如何识别 PCIe 插槽，请参考系统说明文件。
3. 从选择的插槽卸下空挡板。
4. 将适配器的连接器边缘与系统中的 PCIe 连接器插槽对齐。

5. 在适配器卡的两个边角均匀施压以推进插卡，直至其牢固就位在插槽中。当适配器正确就位时，适配器端口连接器将与插槽开口处对齐，适配器面板将与系统机箱齐平。

小心

将插卡推进到位时不要过度用力，否则可能损坏系统或适配器。如果无法固定适配器，将其卸下，重新对齐，并再次尝试。

6. 使用适配器夹或螺丝固定适配器。
7. 合上系统机箱，并断开任何个人防静电设备。

3 驱动程序安装

本章提供有关驱动程序安装的以下信息：

安装 Linux 驱动程序软件

本节介绍如何安装包含或不包含远程直接内存访问 (RDMA) 的 Linux 驱动程序。还介绍 Linux 驱动程序的可选参数、默认值、消息、统计信息以及用于安全引导的公钥。

Dell 支持页面上提供了 41*xxx* 系列适配器 Linux 驱动程序和支持说明文件：

dell.support.com

表 3-1 介绍了 41*xxx* 系列适配器 Linux 驱动程序。

表 3-1. 41xxx 系列适配器 Linux 驱动程序

Linux 驱动程序	说明
qed	qed 核心驱动程序模块直接控制固件，处理中断，并为协议特定驱动动程序集提供低级 API。qed 与 qede、qedr、qedi 和 qedf 驱动程序接合。Linux 核心模块管理所有 PCI 设备资源 （寄存器、主机接口队列，等等）。 qed 核心模块需要使用 Linux 内核版本 2.6.32 或更高版本。测试集中于 x86_64 体系结构。
qede	适用于 41*xxx* 系列适配器的 Linux 以太网驱动程序。该驱动程序直接控制硬件，并负责代表 Linux 主机网络堆栈发送和接收以太网数据包。该驱动程序还代表其自身接收和处理设备中断 （适用于 L2 网络）。 qede 驱动程序需要使用 Linux 内核版本 2.6.32 或更高版本。测试集中于 x86_64 体系结构。
qedr	Linux RoCE 驱动程序在开放结构企业分布 (OFED™) 环境中与 qed 核心模块及 qede 以太网驱动程序联合工作。 RDMA 用户空间应用程序还要求在服务器上安装 libqedr 用户库。
qedi	适用于 41*xxx* 系列适配器的 Linux iSCSI 卸载驱动程序。此驱动程序与 Open iSCSI 库 .y 一起使用。
qedf	适用于 41*xxx* 系列适配器的 Linux FCoE 卸载驱动程序。此驱动程序与 Open FCoE 库一起使用。

可以使用源 Red Hat® Package Manager (RPM) 包或 kmod RPM 包安装 Linux 驱动程序。 RHEL RPM 包内容如下：

- `qlgc-fastlinq-<version>.<OS>.src.rpm`
- `qlgc-fastlinq-kmp-default-<version>.<arch>.rpm`

SLES 源和 kmp RPM 包内容如下：

- `qlgc-fastlinq-<version>.<OS>.src.rpm`
- `qlgc-fastlinq-kmp-default-<version>.<OS>.<arch>.rpm`

以下内核模块 (kmod) RPM 会在运行 Xen 虚拟机监控程序的 SLES 主机上安装 Linux 驱动程序：

- `qlgc-fastlinq-kmp-xen-<version>.<OS>.<arch>.rpm`

以下源 RPM 会在 RHEL 和 SLES 主机上安装 RDMA 库代码：

- `qlgc-libqedr-<version>.<OS>.<arch>.src.rpm`

以下源代码 TAR BZip2 (BZ2) 压缩文件会在 RHEL 和 SLES 主机上安装 Linux 驱动程序：

- `fastlinq-<version>.tar.bz2`

注

要通过 NFS、FTP 或 HTTP（使用网络引导磁盘）进行网络安装，可能需要含 qede 驱动程序的驱动程序磁盘。可以通过修改 makefile 和 make 环境来编译 Linux 引导驱动程序。

安装不含 RDMA 的 Linux 驱动程序

要安装不含 RDMA 的 Linux 驱动程序：

1. 从 Dell 下载 41*xxx* 系列适配器 Linux 驱动程序：

   dell.support.com

2. 如第 10 页上 “ 移除 Linux 驱动程序 ” 中所述，移除现有 Linux 驱动程序。
3. 使用以下方法之一安装新 Linux 驱动程序：
   - ❑ 使用源代码 RPM 软件包安装 Linux 驱动程序
   - ❑ 使用 kmp/kmod RPM 软件包安装 Linux 驱动程序
   - ❑ 使用 TAR 文件安装 Linux 驱动程序

移除 Linux 驱动程序

移除 Linux 驱动程序有两种步骤：一种用于非 RDMA 环境，另一种用于 RDMA 环境。选择符合您的环境的步骤。

要在非 RDMA 环境中移除 Linux 驱动程序，请卸载并移除驱动程序：

按照与原始安装方法和 OS 相关的步骤进行操作。

- 如果以前是使用 RPM 软件包安装的 Linux 驱动程序，请发出以下命令：

```
rmmod qede
rmmod qed
depmod -a
rpm -e qlgc-fastlinq-kmp-default-<version>.<arch>
```

- 如果以前是使用 TAR 文件安装的 Linux 驱动程序，请发出以下命令：

```
rmmod qede
rmmod qed
depmod -a
```

- 对于 RHEL：

```
cd /lib/modules/<version>/extra/qlgc-fastlinq
rm -rf qed.ko qede.ko qedr.ko
```

- 对于 SLES：

```
cd /lib/modules/<version>/updates/qlgc-fastlinq
rm -rf qed.ko qede.ko qedr.ko
```

要在非 RDMA 环境中移除 Linux 驱动程序：

1. 要获取当前安装的驱动程序的路径，请发出以下命令：

```
modinfo <driver name>
```

2. 卸载并移除 Linux 驱动程序。

- 如果以前是使用 RPM 软件包安装的 Linux 驱动程序，请发出以下命令：

```
modprobe -r qede
depmod -a
rpm -e qlgc-fastlinq-kmp-default-<version>.<arch>
```

- 如果以前是使用 TAR 文件安装的 Linux 驱动程序，请发出以下命令：

```
modprobe -r qede
depmod -a
```

注

如果存在 qedr，请发出 `modprobe -r qedr` 命令代替。

3. 从 `qed.ko`、`qede.ko` 和 `qedr.ko` 文件所在的目录中删除它们。例如，在 SLES 中发出以下命令：

```
cd /lib/modules/<version>/updates/qlgc-fastlinq
rm -rf qed.ko
rm -rf qede.ko
rm -rf qedr.ko
depmod -a
```

要在 RDMA 环境中移除 Linux 驱动程序：

1. 要获取安装的驱动程序的路径，请发出以下命令：

```
modinfo <driver name>
```

2. 卸载并移除 Linux 驱动程序。

```
modprobe -r qedr
modprobe -r qede
modprobe -r qed
depmod -a
```

3. 移除驱动程序模块文件：

- ❑ 如果以前是使用 RPM 软件包安装的驱动程序，请发出以下命令：

```
rpm -e qlgc-fastlinq-kmp-default-<version>.<arch>
```

- ❑ 如果以前是使用 TAR 文件安装的驱动程序，则针对您的操作系统发出以下命令：

  对于 RHEL：

```
cd /lib/modules/<version>/extra/qlgc-fastlinq
rm -rf qed.ko qede.ko qedr.ko
```

  对于 SLES：

```
cd /lib/modules/<version>/updates/qlgc-fastlinq
rm -rf qed.ko qede.ko qedr.ko
```

使用源代码 RPM 软件包安装 Linux 驱动程序

要使用源代码 RPM 软件包安装 Linux 驱动程序：

1. 在命令提示符处发出以下命令：

```
rpm -ivh RPMS/<arch>/qlgc-fastlinq-<version>.src.rpm
```

2. 将目录更改至 RPM 路径并针对内核构建二进制 RPM。

注

对于 RHEL 8，在创建二进制 RPM 驱动程序包之前安装 `kernel-rpm-nacros` 和 `kernel-abi-whitelists` 包。

对于 RHEL：

```
cd /root/rpmbuild
rpmbuild -bb SPECS/fastlinq-<version>.spec
```

对于 SLES：

```
cd /usr/src/packages
rpmbuild -bb SPECS/fastlinq-<version>.spec
```

3. 安装新编译的 RPM：

```
rpm -ivh RPMS/<arch>/qlgc-fastlinq-<version>.<arch>.rpm
```

注

如果报告了冲突，则可能需要对某些 Linux 分发版使用 `--force` 选项。

驱动程序将安装在以下路径中。

对于 SLES：

```
/lib/modules/<version>/updates/qlgc-fastlinq
```

对于 RHEL：

```
/lib/modules/<version>/extra/qlgc-fastlinq
```

4. 开启所有 ethX 接口，如下所示：

```
ifconfig <ethX> up
```

5. 对于 SLES，使用 YaST 来配置以太网接口以在引导时自动启动（通过设置静态 IP 地址或启用接口上的 DHCP 实现）。

使用 kmp/kmod RPM 软件包安装 Linux 驱动程序

要安装 kmod RPM 软件包：

1. 在命令提示符处发出以下命令：

```
rpm -ivh qlgc-fastlinq-<version>.<arch>.rpm
```

2. 重新加载驱动程序：

```
modprobe -r qede
modprobe qede
```

使用 TAR 文件安装 Linux 驱动程序

要使用 TAR 文件安装 Linux 驱动程序：

1. 创建目录并将 TAR 文件解压缩到该目录：

```
tar xjvf fastlinq-<version>.tar.bz2
```

2. 切换到最近创建的目录，然后安装驱动程序：

```
cd fastlinq-<version>
make clean; make install
```

qed 和 qede 驱动程序将安装在以下路径中。

对于 SLES：

```
/lib/modules/<version>/updates/qlgc-fastlinq
```

对于 RHEL：

```
/lib/modules/<version>/extra/qlgc-fastlinq
```

3. 加载驱动程序进行测试 （如有必要，先卸载现有驱动程序）：

```
rmmod qede
rmmod qed
modprobe qed
modprobe qede
```

安装包含 RDMA 的 Linux 驱动程序

有关 iWARP 的信息，请参阅第 8 章 iWARP 配置。

要在内建 OFED 环境中安装 Linux 驱动程序：

1. 从 Dell 下载 41*xxx* 系列适配器 Linux 驱动程序：

   dell.support.com

2. 如第 149 页上 “ 在 Linux 的适配器上配置 RoCE” 中所述，在适配器上配置 RoCE。

3. 如第 10 页上 “ 移除 Linux 驱动程序 ” 中所述，移除现有 Linux 驱动程序。

4. 使用以下方法之一安装新 Linux 驱动程序：

   - ❑ 使用 kmp/kmod RPM 软件包安装 Linux 驱动程序
   - ❑ 使用 TAR 文件安装 Linux 驱动程序

5. 安装 libqedr 库以使用 RDMA 用户空间应用程序。 libqedr RPM 仅可用于内建 OFED。在固件支持共存的 RoCE+iWARP 功能之前，必须选择在 UEFI 中使用哪个 RDMA （RoCE、 RoCEv2 或 iWARP）。默认不启用。发出以下命令：

```
rpm –ivh qlgc-libqedr-<version>.<arch>.rpm
```

6. 要创建并安装 libqedr 用户空间库，请发出以下命令：

```
'make libqedr_install'
```

7. 通过加载驱动程序来测试它们，如下所示：

```
modprobe qedr
make install_libeqdr
```

Linux 驱动程序可选参数

表 3-2 介绍 qede 驱动程序的可选参数。

表 3-2. qede 驱动程序可选参数

参数	说明
debug	控制驱动程序详细级别，与 `ethtool -s <dev> msglvl` 类似。
int_mode	控制除 MSI-X 以外的中断模式。
gro_enable	启用或禁用硬件通用接收卸载 (GRO) 功能。此功能与内核的软件 GRO 类似，但只能通过设备硬件执行。
err_flags_override	在发生硬件错误情况下用于禁用或强制采取措施的位图： ■ 位 #31 - 此位掩码的启用位 ■ 位 #0 - 阻止重新断言硬件关注 ■ 位 #1 - 收集调试数据 ■ 位 #2 - 触发恢复过程 ■ 位 #3 - 调用 WARN 以获取导致错误的流的调用跟踪

Linux 驱动程序操作默认值

表 3-3 列出了 qed 和 qede Linux 驱动程序操作默认值。

表 3-3. Linux 驱动程序操作默认值

操作	qed 驱动程序默认值	qede 驱动程序默认值
Speed （速度）	自动协商并广告速度	自动协商并广告速度
MSI/MSI-X	Enabled （已启用）	Enabled （已启用）
Flow Control （流控制）	—	自动协商并广告 RX 和 TX
MTU	—	1500 （范围为 46 - 9600）
Rx Ring Size （Rx 环大小）	—	1000
Tx Ring Size （Tx 环大小）	—	4078 （范围为 128 - 8191）
Coalesce Rx Microseconds （合并 Rx 微秒）	—	24 （范围为 0 - 255）

表 3-3. Linux 驱动程序操作默认值 (续)

操作	qed 驱动程序默认值	qede 驱动程序默认值
Coalesce Tx Microseconds （合并 Tx 微秒）	—	48
TSO	—	Enabled （已启用）

Linux 驱动程序消息

要设置 Linux 驱动程序消息详细级别，请发出以下命令之一：

- **ethtool -s <interface> msglvl <value>**
- **modprobe qede debug=<value>**

  其中 <value> 表示 0-15 位，这些是标准的 Linux 网络值，并且 16 及更高的位特定于驱动程序。

统计信息

要查看详细的统计信息和配置信息，请使用 ethtool 公用程序。参见 ethtool 手册页了解更多信息。

导入用于安全引导的公钥

Linux 驱动程序要求您导入并注册 Qlogic 公钥以在安全引导环境中加载驱动程序。开始之前，确保您的服务器支持安全引导。本节提供两种导入并注册公钥的方法。

要导入并注册 Qlogic 公钥：

1. 从以下网页下载公钥：

   http://ldriver.qlogic.com/Module-public-key/

2. 要安装公钥，请发出以下命令：

   # mokutil --root-pw --import cert.der

   其中 --root-pw 选项用于根用户的直接使用。

3. 重新引导系统。
4. 检查准备注册的证书列表：

   # mokutil --list-new

5. 再次重新引导系统。
6. 当 shim 启动 MokManager 时，输入根密码以确认证书导入到 Machine Owner Key (MOK) 列表。

7. 要确定新导入的密钥是否已注册：

```
# mokutil --list-enrolled
```

要手动启动 MOK 并注册 Qlogic 公钥：

1. 发出以下命令：

```
# reboot
```

2. 在 **GRUB 2** 菜单中，按 C 键。
3. 发出以下命令：

```
chainloader $efibootdir/MokManager.efi
- boot
```

4. 选择 **Enroll key from disk** （从磁盘注册密钥）。
5. 导航到 `cert.der` 文件，然后按 ENTER 键。
6. 按照说明注册密钥。一般包括按下 0 （零）键，然后按 Y 键确认。

注

固件菜单可能提供更多新增密钥到签名数据库的方法。

有关安全引导的其他信息，请参阅以下网页：

https://www.suse.com/documentation/sled-12/book_sle_admin/data/sec_uefi_secboot.html

安装 Windows 驱动程序软件

有关 iWARP 的信息，请参阅第 8 章 iWARP 配置。

- 安装 Windows 驱动程序
- 移除 Windows 驱动程序
- 管理适配器属性
- 设置电源管理选项
- Windows 中的链路配置

安装 Windows 驱动程序

使用 Dell Update Package (DUP) 安装 Windows 驱动程序软件：

- 在 GUI 中运行 DUP
- DUP 安装选项
- DUP 安装示例

在 GUI 中运行 DUP

要在 GUI 中运行 DUP：

1. 双击代表 Dell Update Package 文件的图标。

注

Dell Update Package 的实际文件名称将会不同。

2. 在 Dell Update Package 窗口 （图 3-1）中，单击 **Install** （安装）。

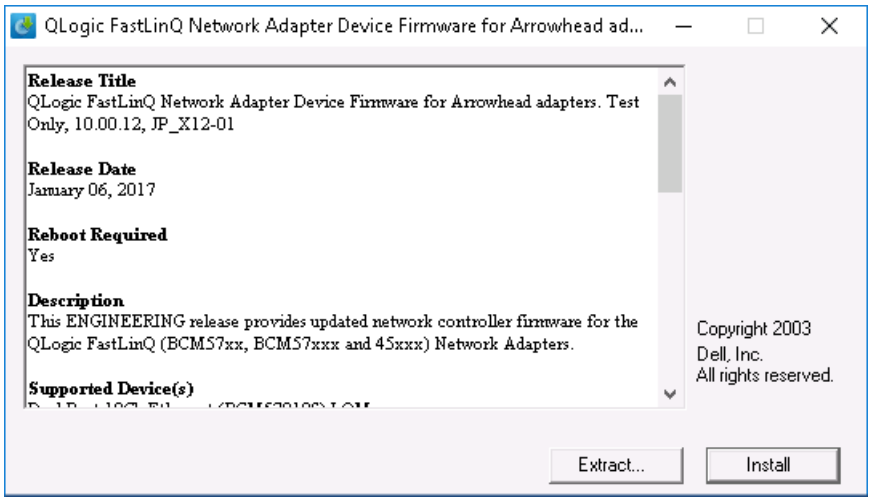


图 3-1. Dell Update Package 窗口

3. 在 QLogic Super Installer—InstallShield® 向导的 Welcome （欢迎）窗口（图 3-2）中，单击 **Next** （下一步）。

图 3-2. QLogic InstallShield 向导：Welcome （欢迎）窗口

4. 在向导的 License Agreement （许可协议）窗口 （图 3-3）中完成以下操作：
   a. 阅读 End User Software License Agreement （最终用户许可协议）。
   b. 选择 **I accept the terms in the license agreement** （我接受许可协议中的条款）继续。
   c. 单击 **Next** （下一步）。

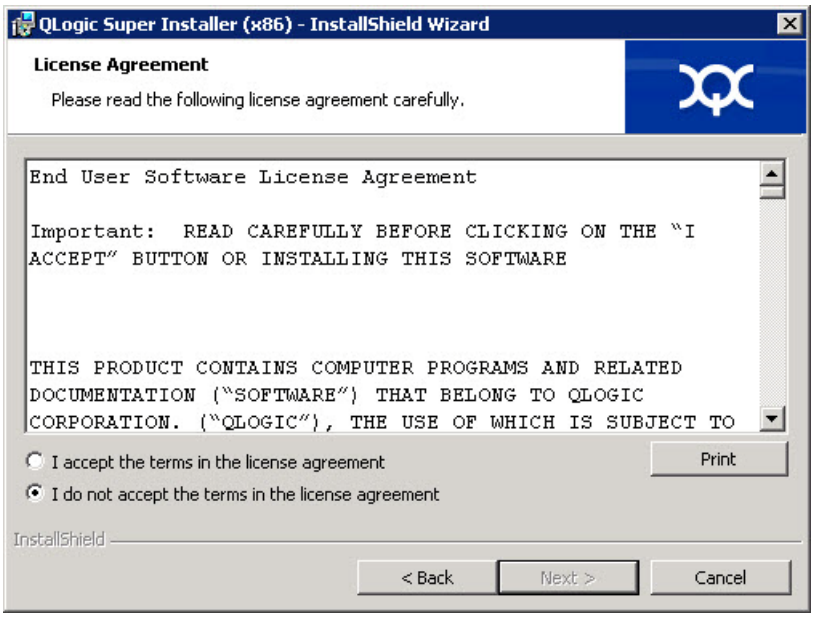


图 3-3. QLogic InstallShield 向导：License Agreement （许可协议）窗口

5. 如下完成向导的 Setup Type （设置类型）窗口 （图 3-4）：

   a. 选择以下设置类型之一：

      - 单击 **Complete** （完成）以安装所有程序功能。
      - 单击 **Custom** （自定义）以手动选择要安装的功能。

   b. 单击 **Next** （下一步）继续。

   如果单击 **Complete** （完成），则直接继续步骤 6b。

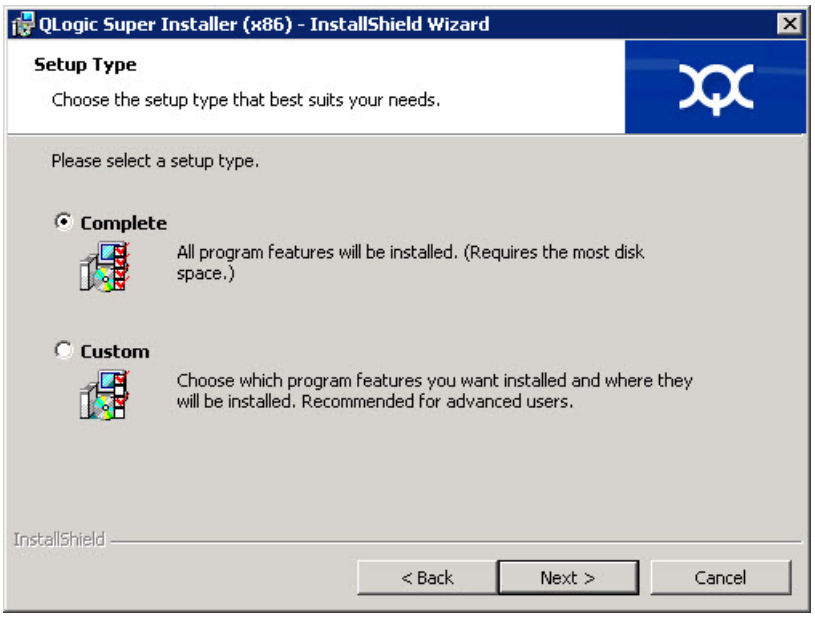


图 3-4. InstallShield 向导：Setup Type （设置类型）窗口

6. 如果您在步骤 5 中选择 **Custom** （自定义），则如下完成 Custom Setup （自定义设置）窗口 （图 3-5）：

   a. 选择要安装的功能。默认情况下，将选中所有功能。要更改某个功能的安装设置，请单击该功能旁边的图标，然后选择以下选项之一：

      - **This feature will be installed on the local hard drive** （此功能将安装在本地硬盘驱动器上）- 标记安装的功能，但不会影响其任何子功能。
      - **This feature， and all subfeatures， will be installed on the local hard drive** （此功能及所有子功能都将安装在本地硬盘驱动器上） - 标记安装的功能及其所有子功能。
      - **This feature will not be available** （此功能将不可用） - 阻止安装该功能。

   b. 单击 **Next** （下一步）继续。

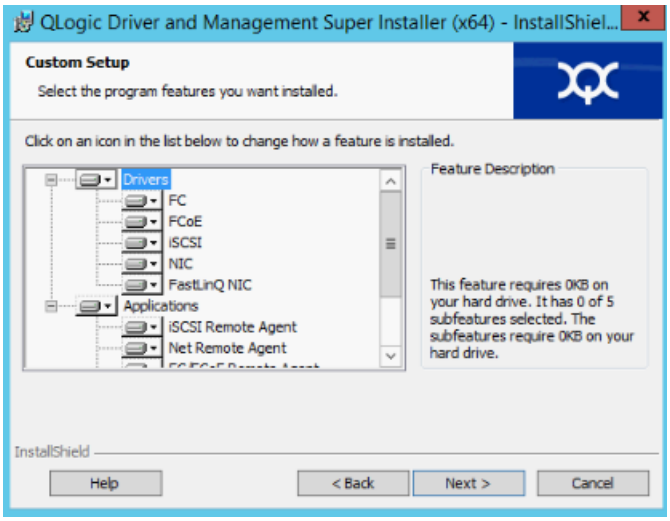


图 3-5. InstallShield 向导：Custom Setup （自定义设置）窗口

7. 在 InstallShield 向导的 Ready To Install （准备安装）窗口 （图 3-6）中，单击 **Install** （安装）。 InstallShield 向导将安装 QLogic 适配器驱动程序和管理软件安装程序。

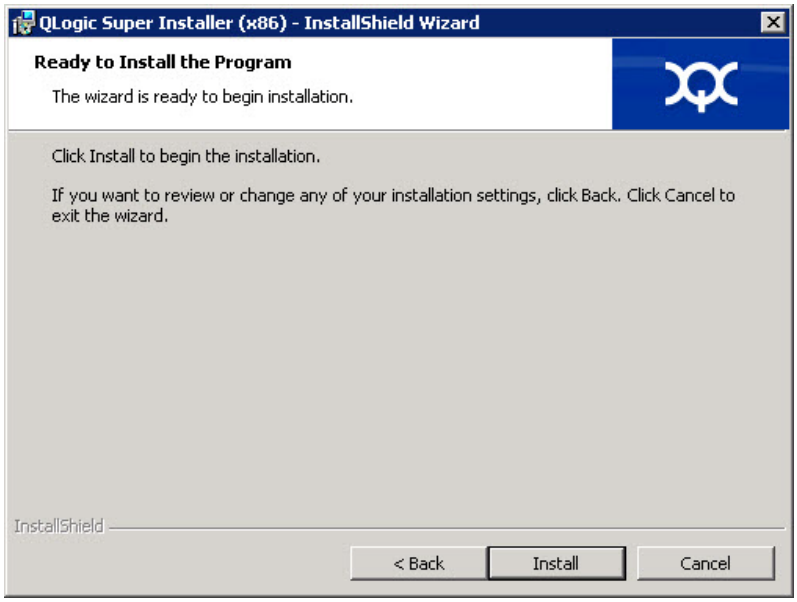


图 3-6. InstallShield 向导：Ready to Install the Program （准备安装程序）窗口

8. 安装完成时，将显示 InstallShield Wizard Completed （InstallShield 向导完成）窗口 （图 3-7）。单击 **Finish** （完成）关闭安装程序。

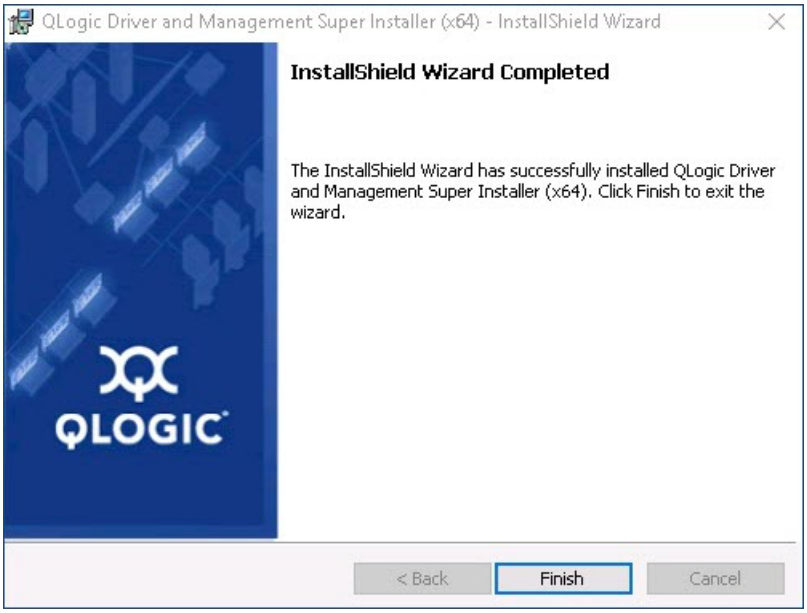


图 3-7. InstallShield 向导：Completed （完成）窗口

9. 在 Dell Update Package 窗口（图 3-8）中，“Update installer operation was successful （更新安装程序操作成功） ” 表示完成。

   - ❑ （可选）要打开日志文件，请单击 **View Installation Log** （查看安装日志）。日志文件显示 DUP 安装进度、任何之前安装的版本、任何错误消息，以及有关安装的其他信息。
   - ❑ 要关闭 Update Package （更新软件包）窗口，请单击 **CLOSE** （关闭）。

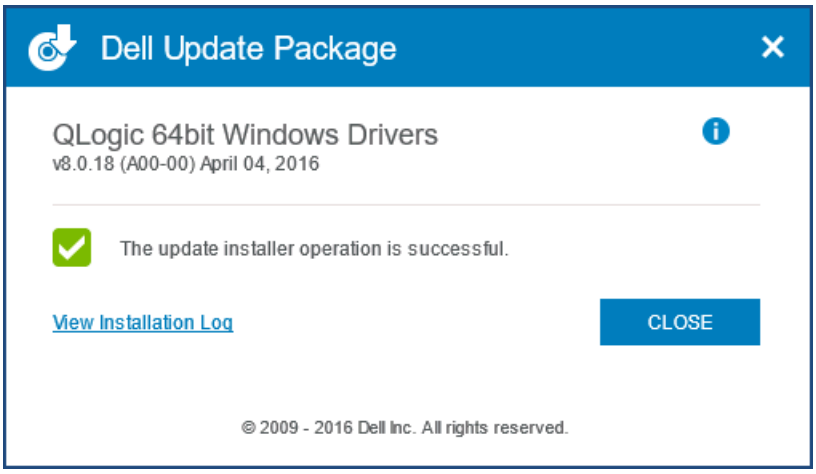


图 3-8. Dell Update Package 窗口

DUP 安装选项

要自定义 DUP 安装行为，请使用以下命令行选项。

- 要仅将驱动程序组件提取到某个目录，请使用以下选项：

  /drivers=<path>

  注

  此命令需要 **/s** 选项。

- 要仅安装或更新驱动程序组件，请使用以下选项：

  /driveronly

  注

  此命令需要 **/s** 选项。

- （高级）使用 /passthrough 选项可将 passthrough 后的所有文本直接发送至 DUP 的 QLogic 安装软件。此模式禁用提供的所有 GUI，但不一定禁用 QLogic 软件的 GUI。

  /passthrough

- （高级）要返回此 DUP 支持的功能的代码说明，请使用以下选项：

  /capabilities

注

此命令需要 **/s** 选项。

DUP 安装示例

以下示例显示如何使用安装选项。

要以无提示方式更新系统，请使用以下命令：

```
<DUP_file_name>.exe /s
```

要将更新内容提取到 C:\mydir\ 目录：

```
<DUP_file_name>.exe /s /e=C:\mydir
```

要将驱动程序组件提取到 C:\mydir\ 目录：

```
<DUP_file_name>.exe /s /drivers=C:\mydir
```

要仅安装驱动程序组件，请使用以下命令：

```
<DUP_file_name>.exe /s /driveronly
```

要从默认日志位置更改为 C:\my path with spaces\log.txt，请使用以下命令：

```
<DUP_file_name>.exe /l="C:\my path with spaces\log.txt"
```

移除 Windows 驱动程序

要移除 Windows 驱动程序：

1. 在控制面板中，单击 **Programs**（程序），然后单击 **Programs and Features**（程序和功能）。
2. 在程序列表中，选择 **QLogic FastLinQ Driver Installer**（QLogic FastLinQ 驱动程序安装程序），然后单击 **Uninstall**（卸载）。
3. 按照说明操作以移除驱动程序。

管理适配器属性

要查看或更改 41*xxx* 系列适配器属性：

1. 在控制面板中，单击 **Device Manager**（设备管理器）。
2. 在所选适配器的属性上，单击 **Advanced**（高级）选项卡。
3. 在 Advanced（高级）页面（图 3-9）上，选择 **Property**（属性）下的项目，然后根据需要更改该项目的 **Value**（值）。

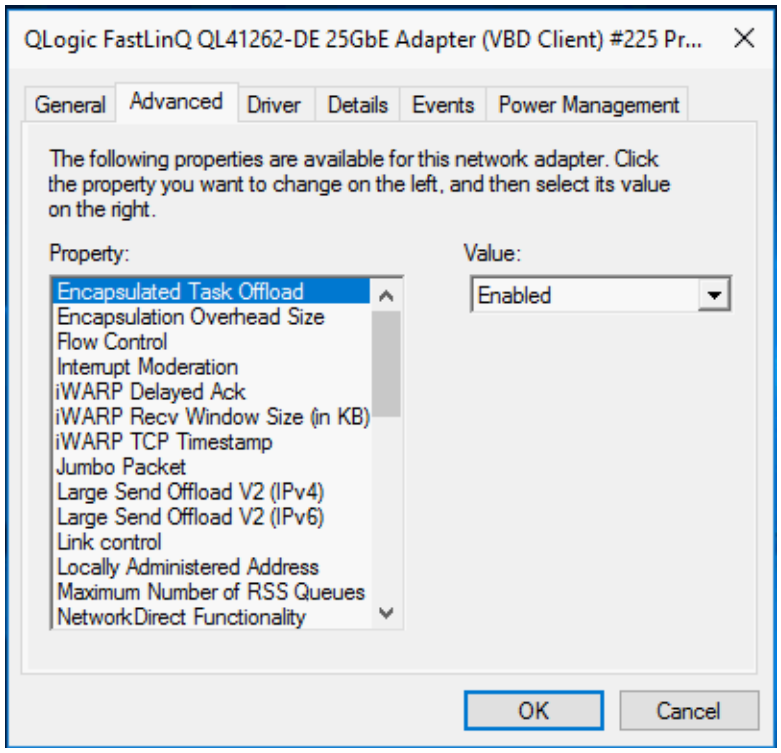


图 3-9. 设置高级适配器属性

设置电源管理选项

可以设置电源管理选项，以允许操作系统关闭该控制器以节约电源，或者允许该控制器唤醒计算机。如果设备正忙 （例如，正在处理呼叫），操作系统将不会关闭设备。只有在计算机试图进入休眠状态时，操作系统才尝试尽可能关闭各个设备。要使控制器一直保持打开状态，不要选择 **Allow the computer to turn off the device to save power** （允许计算机关闭此设备以节约电源）复选框（图 3-10）。

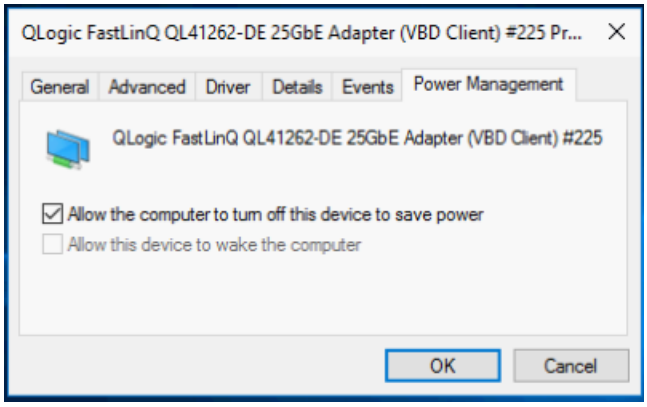


图 3-10. 电源管理选项

注

- 只有支持电源管理的服务器才有 Power Management （电源管理）页面。
- 对于作为组成员的任何适配器，不要选中 **Allow the computer to turn off the device to save power** （允许计算机关闭设备以节约电源）复选框。

配置通信协议以使用 QCC GUI、 QCC PowerKit 和 QCS CLI

QCCGUI、 QCC PowerKit 和 QCS CLI 管理应用程序有两个主要组件：RPC 代理程序和客户端软件。 RPC 代理程序安装在包含一个或多个聚合网络适配器的服务器 （即受管主机）上。 RPC 代理程序搜集聚合网络适配器上的信息，并将其供安装有客户端软件的管理电脑检索。客户端软件可以从 RPC 代理程序查看信息，并且配置聚合网络适配器。管理软件包含 QCC GUI 和 QCS CLI。

通信协议可使 RPC 代理程序与客户端软件之间进行通信。根据网络中客户端和受管主机上混合使用的操作系统 （Linux 和 / 或 Windows），可以选择使用合适的公用程序。

有关这些管理应用程序的安装说明，请参阅 Marvell 网站上的以下文档：

- *QLogic Control Suite CLI 用户指南*（文档号 BC0054511-00）
- *PowerShell 用户指南*（文档号 BC0054518-00）
- *QConvergeConsole GUI 安装指南*（文档号 SN0051105-00）

Windows 中的链路配置

在 Windows OS 中可使用三个不同的参数进行链路配置，可用于 Device Manager（设备管理器）页面中 Advanced （高级）选项卡上的配置。

链路控制模式

控制链路配置有两种模式：

- **Preboot Controlled** （预引导控制）是默认模式。在此模式中，驱动程序使用设备中的链路配置，可从预引导组件配置。此模式会忽略 Advanced （高级）选项卡上的链路参数。
- 要从 Device Manager （设备管理器）页面的 Advanced （高级）选项卡 （如图 3-11 所示）配置链路设置时，应设置 **Driver Controlled** （驱动程序控制）模式。

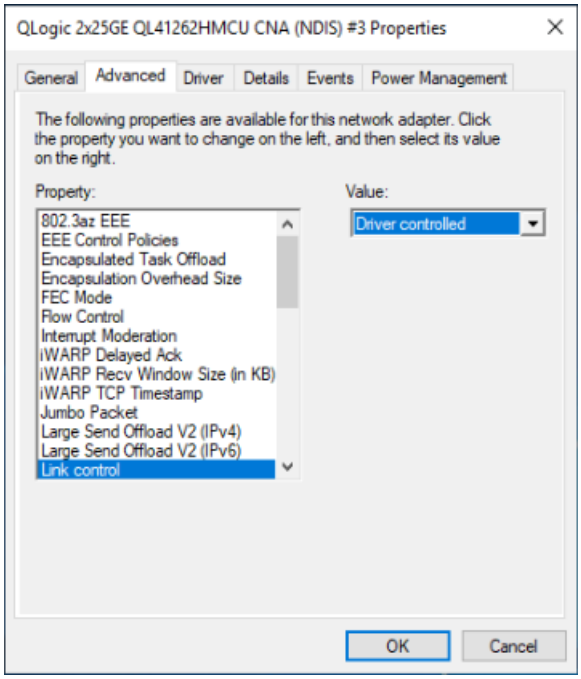


图 3-11. 设置 Driver Controlled （驱动程序控制）模式

链路速度和双工

Speed & Duplex （速度和双工）属性 （在 Device Manager （设备管理器）页面的 Advanced （高级）选项卡上）可配置为 Value （值）菜单 （参见图 3-12）中的任意选择。

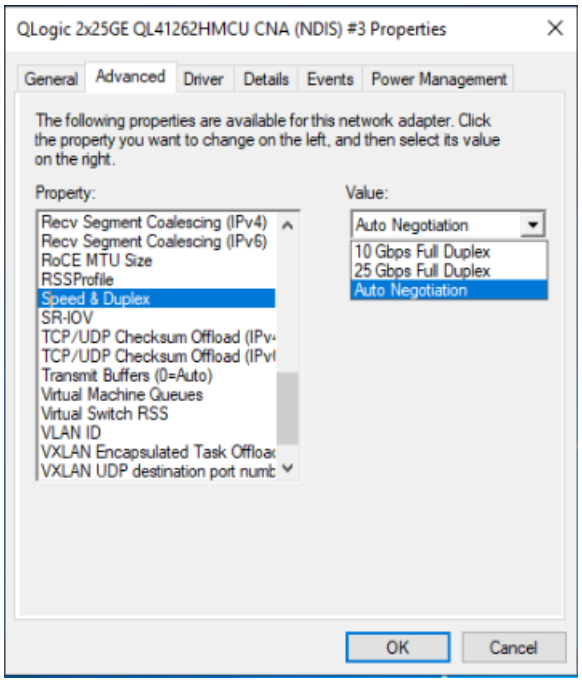


图 3-12. 设置 Link Speed and Duplex （链路速度和双工）属性

此配置仅在链路控制属性设为 Driver controlled （驱动程序控制）（参见图 3-11）时有效。

FEC 模式

OS 级的 FEC 模式配置涉及三种驱动程序高级属性。

要设置 FEC 模式：

1. 设置链路控制。在 Device Manager （设备管理器）页面的 Advanced （高级）选项卡上：
   a. 在 Property （属性）菜单中选择 **Link control** （链路控制）。
   b. 在 Value （值）菜单中选择 **Driver controlled** （驱动程序控制）。

   有关示例请参见图 3-11。

2. 设置速度和双工。在 Device Manager （设备管理器）页面的 Advanced （高级）选项卡上：

   a. 在 Property （属性）菜单中选择 **Speed & Duplex** （速度和双工）。

   b. 在 Value （值）菜单中选择固定速度。

   FEC 模式配置仅在 Speed & Duplex （速度和双工）设为固定速度时才会激活。将此属性设置为 Auto Negotiation （自动协商）会禁用 FEC 配置。

3. 设置 FEC 模式。在 Device Manager （设备管理器）页面的 Advanced （高级）选项卡上：

   a. 在 Property （属性）菜单中选择 **FEC Mode** （FEC 模式）。

   b. 在 Value （值）菜单中选择有效值 （参见图 3-13）。

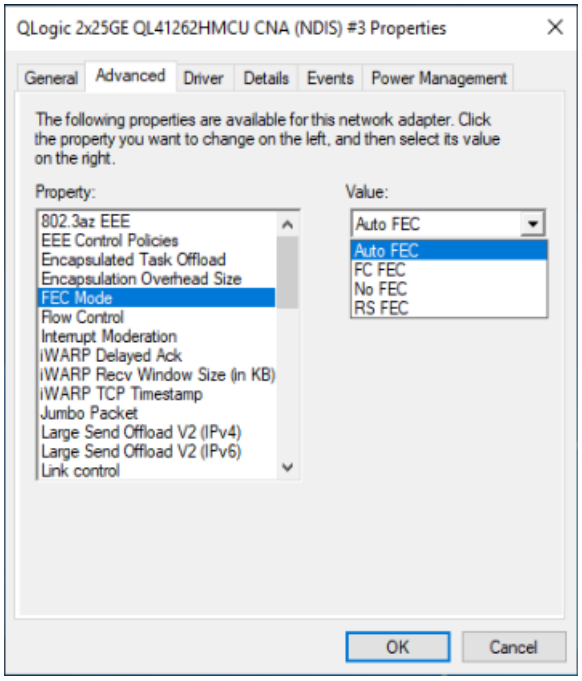


图 3-13. 设置 FEC 模式属性

此属性仅在步骤 1 和步骤 2 完成后才生效。

并非所有 FEC 模式对每一种介质都有效；您必须知道对您的特定介质有效的模式。如果设置了错误的 FEC 模式值，链路将会关闭。

安装 VMware 驱动程序软件

本节介绍用于 41*xxx* 系列适配器的 qedentv VMware ESXi 驱动程序：

VMware 驱动程序和驱动程序包

表 3-4 列出了协议的 VMware ESXi 驱动程序。

表 3-4. VMware 驱动程序

VMware 驱动程序	说明
qedentv	本机网络驱动程序
qedrntv	本机 RDMA 卸载 （RoCE 和 RoCEv2）驱动程序[a]
qedf	本机 FCoE 卸载驱动程序
qedil	旧版 iSCSI 卸载驱动程序
qedi	本机 iSCSI 卸载驱动程序 （ESXi 6.7 及更高版本）[b]

[a] 对于 ESXi 6.5，NIC 和 RoCE 驱动程序已打包在一起，并且可使用标准 ESXi 安装命令作为单个脱机 zip 捆绑包进行安装。建议的安装序列依次是 NIC/RoCE 驱动程序包、 FCoE 和 iSCSI 驱动程序包 （按需要）。

[b] 对于 ESXi 6.7， NIC、 RoCE 及 iSCSI 驱动程序已打包在一起，并且可使用标准 ESXi 安装命令作为单个脱机 zip 捆绑包进行安装。建议的安装序列依次是 NIC/RoCE/iSCSI 驱动程序包、 FCoE 驱动程序包 （按需要）。

ESXi 驱动程序作为单独的驱动程序包随附，除非另有说明，否则不会捆绑在一起。

VMware 驱动程序只能从 VMware 网站下载。

https://www.vmware.com/resources/compatibility/search.php?deviceCategory=io&details=1&keyword=QL41&page=1&display_interval=10&sortColumn=Partner&sortOrder=Asc

使用以下任一方法安装单独的驱动程序：

- 标准 ESXi 软件包安装命令 （请参阅安装 VMware 驱动程序）
- 单独驱动程序自述文件中的步骤
- 以下 VMware KB 文章中的步骤：

  https://kb.vmware.com/selfservice/microsites/search.do?language=en_US&cmd=displayKC&externalId=2137853

您应该首先安装 NIC 驱动程序，然后再安装存储驱动程序。

安装 VMware 驱动程序

可以使用驱动程序 ZIP 文件安装新驱动程序或更新现有驱动程序。请务必从同一驱动程序 ZIP 文件安装整个驱动程序集。混用来自不同 ZIP 文件的驱动程序会导致问题。

要安装 VMware 驱动程序：

1. 从 VMware 支持页面下载适用于 41*xxx* 系列适配器的 VMware 驱动程序：

   www.vmware.com/support.html

2. 启动 ESX 主机，然后登录至具有管理员权限的帐户。
3. 使用 Linux scp 公用程序将驱动程序包从本地系统复制到 IP 地址为 10.10.10.10 的 ESX 服务器上的 `/tmp` 目录中。例如，发出以下命令：

   ```
   # scp qedentv-bundle-2.0.3.zip root@10.10.10.10:/tmp
   ```

   您可以将该文件放在 ESX 控制台 shell 可以访问的任意位置。
4. 通过发出以下命令，将主机置于维护模式：

   ```
   # esxcli --maintenance-mode
   ```

注

ESXi 主机支持的最大 qedentv 以太网接口数为 32，因为 vmkernel 只允许 32 个接口注册管理回呼。

5. 选择以下安装选项之一：

   ❑ **选项 1**：发出以下命令，安装驱动程序包 （将一次安装所有驱动程序 VIB）：

   ```
   # esxcli software vib install -d /tmp/qedentv-2.0.3.zip
   ```

   ❑ **选项 2**：使用 CLI 或 VMware Update Manager (VUM) 直接在 ESX 服务器上安装 .vib。为此，请解压缩驱动程序 ZIP 文件，然后提取 .vib 文件。

   ■ 要使用 CLI 安装 .vib 文件，请发出以下命令。请务必指定完整的 .vib 文件路径：

```
# esxcli software vib install -v /tmp/qedentv-1.0.3.11-1OEM.550.0.0.1331820.x86_64.vib
```

   ■ 要使用 VUM 安装 .vib 文件，请参阅下面的知识库文章：

   使用 VMware vCenter Update Manager 4.x 和 5.x 更新 ESXi/ESX 主机 (1019545)

要升级现有的驱动程序包：

■ 发出以下命令：

```
# esxcli software vib update -d /tmp/qedentv-bundle-2.0.3.zip
```

要升级个别驱动程序：

执行全新安装的步骤 （请参阅**要安装 VMware 驱动程序**），只是将 “ 选项 1” 中的命令替换成以下命令：

```
# esxcli software vib update -v /tmp/qedentv-1.0.3.11-1OEM.550.0.0.1331820.x86_64.vib
```

VMware NIC 驱动程序可选参数

表 3-5 说明了可作为命令行参数提供给 esxcfg-module 命令的可选参数。

表 3-5. VMware NIC 驱动程序可选参数

参数	说明
hw_vlan	全局启用 (1) 或禁用 (0) 硬件 vLAN 插入和移除。当上层需要发送或接收完整格式数据包时，禁用此参数。 hw_vlan=1 为默认值。
num_queues	指定 TX/RX 队列对的数目。 num_queues 可以是 1-11 或以下值之一： ■ -1 允许驱动程序决定队列对的最佳数目 （默认值）。 ■ 0 使用默认队列。 对于多端口或多功能配置，您可以指定用逗号分隔的多个值。

表 3-5. VMware NIC 驱动程序可选参数 (续)

参数	说明
multi_rx_filters	指定每个 RX 队列的 RX 过滤器数目，默认队列除外。 multi_rx_filters 可以是 1-4 或以下值之一： ■ -1 使用默认的每队列 RX 过滤器数目。 ■ 0 禁用 RX 过滤器。
disable_tpa	启用 (0) 或禁用 (1) TPA (LRO) 功能。 disable_tpa=0 为默认值。
max_vfs	指定每个物理功能 (PF) 的虚拟功能 (VF) 数目。 max_vfs 可以是一个端口上的 0 个（已禁用）或 64 个 VF （已启用）。 ESXi 支持最大 64 个 VF 是 OS 资源分配约束。
RSS	指定主机或 PF 的虚拟可扩展 LAN (VXLAN) 隧道流量所使用的接收端伸缩队列数目。 RSS 可以是 2、 3、 4 或以下值之一： ■ -1 使用默认队列数目。 ■ 0 或 1 禁用 RSS 队列。 对于多端口或多功能配置，您可以指定用逗号分隔的多个值。
debug	指定驱动程序在 vmkernel 日志文件中记录的数据级别。 debug 可具有以下值 （按数据量升序显示）： ■ 0x80000000 表示通知级别。 ■ 0x40000000 表示信息级别 （包括通知级别）。 ■ 0x3FFFFFFF 表示针对所有驱动程序子模块的详细级别 （包括信息和通知级别）。
auto_fw_reset	启用 (1) 或禁用 (0) 驱动程序自动固件恢复功能。启用此参数时，驱动程序会尝试从发送超时、固件断言和适配器奇偶校验错误等事件中恢复。默认值为 auto_fw_reset=1。
vxlan_filter_en	启用 (1) 或禁用 (0) 基于外层 MAC、内层 MAC 和 VXLAN 网络 (VNI) 的 VXLAN 过滤，将流量直接匹配至特定队列。默认值为 vxlan_filter_en=1。对于多端口或多功能配置，您可以指定用逗号分隔的多个值。
enable_vxlan_offld	启用 (1) 或禁用 (0) VXLAN 隧道流量校验和卸载和 TCP 分段卸载 (TSO) 功能。默认值为 enable_vxlan_offld=1。对于多端口或多功能配置，您可以指定用逗号分隔的多个值。

VMware 驱动程序参数默认值

表 3-6 列出了 VMware 驱动程序参数默认值。

表 3-6. VMware 驱动程序参数默认值

参数	默认值
Speed （速度）	自动协商并广告所有速度。 speed （速度）参数在所有端口上必须相同。如果在设备上启用了自动协商，则所有设备端口都将使用自动协商。
Flow Control （流控制）	自动协商并广告 RX 和 TX
MTU	1,500 （范围为 46 - 9,600）
Rx Ring Size （Rx 环大小）	8,192 （范围为 128 - 8,192）
Tx Ring Size （Tx 环大小）	8,192 （范围为 128 - 8,192）
MSI-X	Enabled （已启用）
Transmit Send Offload （传输发送卸载） (TSO)	Enabled （已启用）
Large Receive Offload （大量接收卸载） (LRO)	Enabled （已启用）
RSS	Enabled （已启用）（四个 RX 队列）
HW VLAN	Enabled （已启用）
Number of Queues （队列数）	Enabled （已启用）（八个 RX/TX 队列对）
Wake on LAN （局域网唤醒） (WoL)	Disabled （已禁用）

移除 VMware 驱动程序

要移除 .vib 文件 (qedentv)，请发出以下命令：

```
# esxcli software vib remove --vibname qedentv
```

要移除驱动程序，请发出以下命令：

```
# vmkload_mod -u qedentv
```

FCoE 支持

Marvell VMware FCoE qedf 驱动程序包含在 VMware 软件包中，用于支持 Marvell FastLinQ FCoE 聚合网络接口控制器 (C-NIC)。该驱动程序是内核模式驱动程序，提供 VMware SCSI 堆栈与 Marvell FCoE 固件和硬件之间的转换层。VMware ESXi 5.0 及以上版本支持 FCoE 和 DCB 功能。

要启用 FCoE 卸载模式，请参阅位于 https://www.marvell.com/documents/5aa5otcbkr0im3ynera3/ 的 *Application Note, Enabling Storage Offloads on Dell and Marvell FastLinQ 41000 Series Adapters*。

iSCSI 支持

Marvell VMware iSCSI qedil 主机总线适配器 (HBA) 驱动程序类似于 qedf，是内核模式驱动程序，提供 VMware SCSI 堆栈与 Marvell iSCSI 固件和硬件之间的转换层。 qedil 驱动程序利用 VMware iscsid 基础设施用于会话管理和 IP 服务的服务。

要启用 iSCSI 卸载模式，请参阅位于 https://www.marvell.com/documents/5aa5otcbkr0im3ynera3/ 的 *Application Note, Enabling Storage Offloads on Dell and Marvell FastLinQ 41000 Series Adapters*。

注

QL41xxx*xxx* 适配器支持的 iSCSI 接口是相关的硬件接口，依赖网络服务、iSCSI 配置和 VMware 提供的管理接口。 iSCSI 接口包含两个组件：同一接口上的网络适配器和 iSCSI 引擎。 iSCSI 引擎在存储适配器列表上显示为 iSCSI 适配器 (vmhba)。对于 iSCSI 需要的 ARP 和 DHCP 等服务， iSCSI vmhba 使用 qedil 驱动程序创建的 vmnic 设备的服务。 vmnic 是一种瘦虚拟实现，用于向 iSCSI 提供 L2 功能进行操作。不要配置、分配到虚拟交换机或以用于传输常规网络流量的任何方式使用 vmnic。适配器上的实际 NIC 接口将由 qedentv 驱动程序 （全功能 NIC 驱动程序）声明。

4 升级固件

本章提供有关使用 Dell Update Package (DUP) 升级固件的信息。

固件 DUP 是仅用于更新闪存的公用程序；它不用于适配器配置。您可通过双击可执行文件来运行固件 DUP。或者，您可使用多个支持的命令行选项从命令行运行 DUP。

通过双击运行 DUP

要通过双击可执行文件来运行固件 DUP：

1. 双击代表固件 Dell Update Package 文件的图标。
2. 随即出现 Dell Update Package 初始屏幕，如图 4-1 中所示。单击 **Install**（安装）以继续。

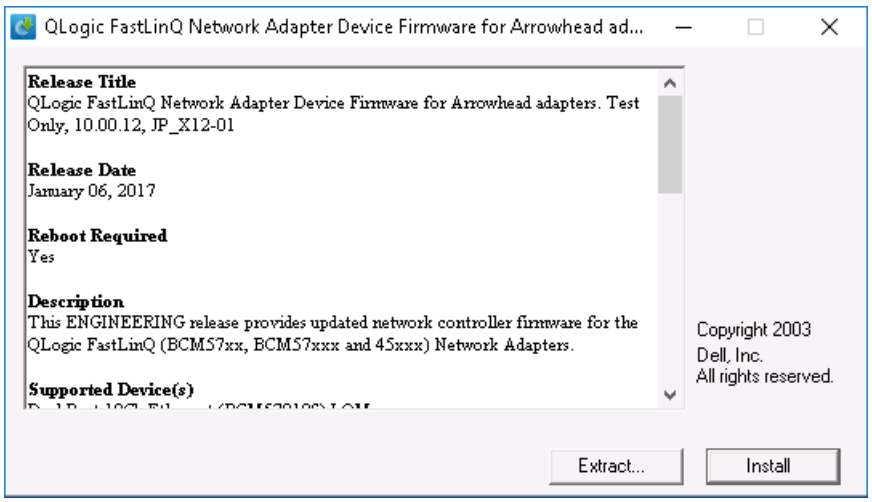


图 4-1. Dell Update Package：初始屏幕

3. 请按照屏幕说明进行操作。在 Warning （警告）对话框中，单击 **Yes** （是）继续安装。

   安装程序表示正在加载新固件，如 图 4-2 所示。

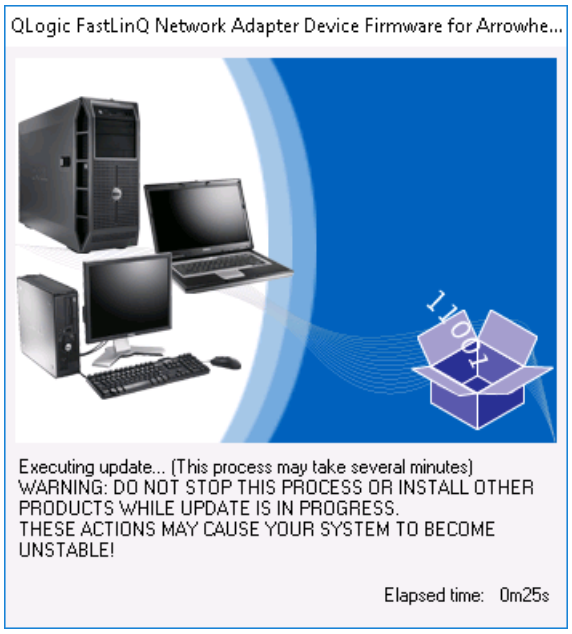


图 4-2. Dell Update Package：加载新固件

完成时，安装程序表明安装结果，如 图 4-3 所示。

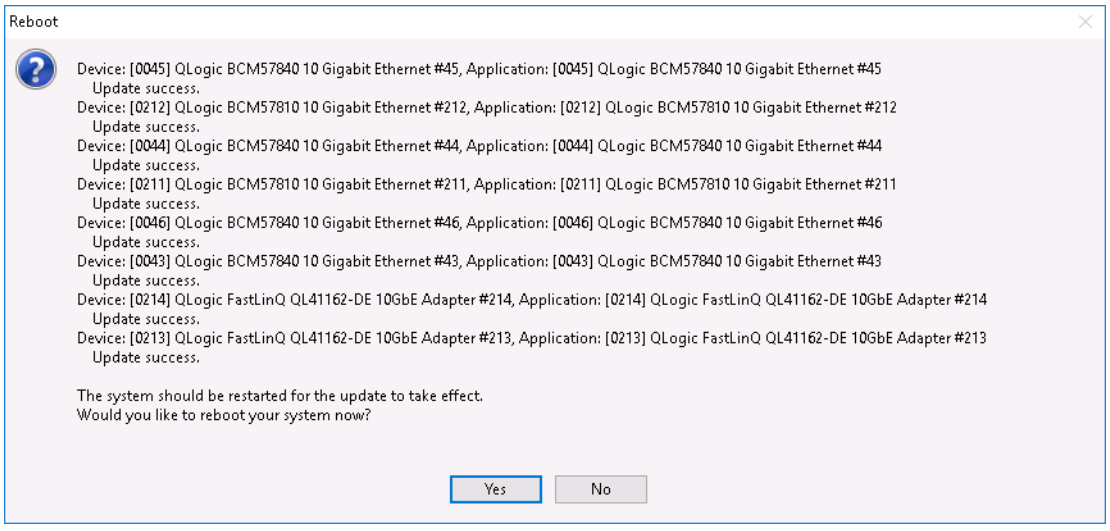


图 4-3. Dell Update Package：安装结果

4. 单击 **Yes** （是）重新引导系统。
5. 单击 **Finish** （完成）以完成安装，如 图 4-4 所示。

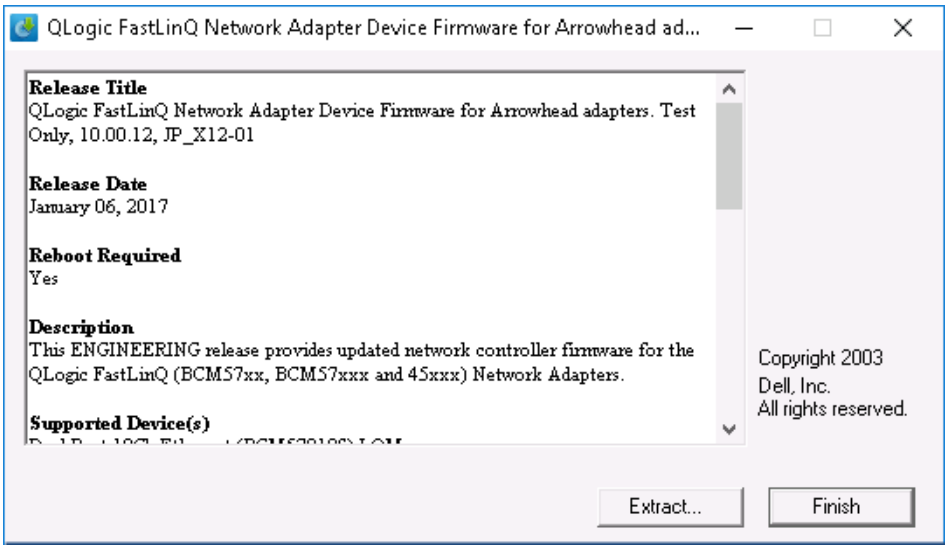


图 4-4. Dell Update Package：完成安装

从命令行运行 DUP

从命令行运行固件 DUP 而不指定选项时，会产生与双击 DUP 图标相同的行为。请注意， DUP 的实际文件名称各有不同。

要从命令行运行固件 DUP：

- 发出以下命令：

```
C:\> Network_Firmware_2T12N_WN32_<version>_X16.EXE
```

图 4-5 显示可用于自定义 Dell Update Package 安装的选项。

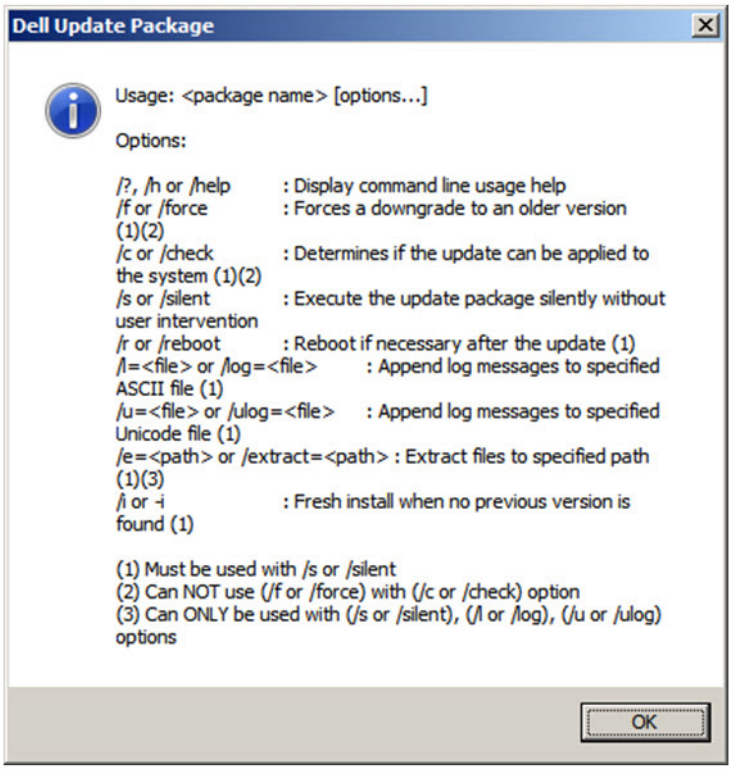


图 4-5. DUP 命令行选项

使用 .bin 文件运行 DUP

以下步骤仅在 Linux OS 上受支持。

要使用 .bin 文件更新 DUP：

1. 将 `Network_Firmware_NJCX1_LN_X.Y.Z.BIN` 文件复制到系统或服务器。
2. 将文件类型更改为可执行文件，如下所示：

   ```
   chmod 777 Network_Firmware_NJCX1_LN_X.Y.Z.BIN
   ```

3. 要开始更新过程，请发出以下命令：

   ```
   ./Network_Firmware_NJCX1_LN_X.Y.Z.BIN
   ```

4. 固件更新后，重新引导系统。

DUP 更新期间 SUT 的示例输出：

```
./Network_Firmware_NJCX1_LN_08.07.26.BIN
Collecting inventory...
Running validation...
BCM57810 10 Gigabit Ethernet rev 10 (p2p1)
The version of this Update Package is the same as the currently installed
version.
Software application name: BCM57810 10 Gigabit Ethernet rev 10 (p2p1)
Package version: 08.07.26
Installed version: 08.07.26
BCM57810 10 Gigabit Ethernet rev 10 (p2p2)
The version of this Update Package is the same as the currently installed
version.
Software application name: BCM57810 10 Gigabit Ethernet rev 10 (p2p2)
Package version: 08.07.26
Installed version: 08.07.26
Continue? Y/N:Y
Y entered; update was forced by user
Executing update...
WARNING: DO NOT STOP THIS PROCESS OR INSTALL OTHER DELL PRODUCTS WHILE UPDATE
IS IN PROGRESS.
THESE ACTIONS MAY CAUSE YOUR SYSTEM TO BECOME UNSTABLE!
.....................................................................
Device: BCM57810 10 Gigabit Ethernet rev 10 (p2p1)
  Application: BCM57810 10 Gigabit Ethernet rev 10 (p2p1)
  Update success.
Device: BCM57810 10 Gigabit Ethernet rev 10 (p2p2)
  Application: BCM57810 10 Gigabit Ethernet rev 10 (p2p2)
  Update success.
Would you like to reboot your system now?
Continue? Y/N:Y
```

5 适配器预引导配置

在主机引导过程中，您有机会暂停并使用人机界面基础设施 (HII) 应用程序执行适配器管理任务。包括以下任务：

注

本章中的 HII 屏幕截图具有代表性，可能不符合您在系统上看到的屏幕。

启动

要启动 HII 应用程序：

1. 打开您的平台的 System Setup （系统设置）窗口。有关启动 System Setup （系统设置）的信息，请查询您的系统的用户指南。
2. 在 System Setup （系统设置）窗口 （图 5-1）中，选择 **Device Settings** （设备设置），然后按 ENTER 键。

图 5-1. 系统设置

3. 在 Device Settings （设备设置）窗口 （图 5-2）中，选择您要配置的 41*xxx* 系列适配器 端口，然后按 ENTER 键。

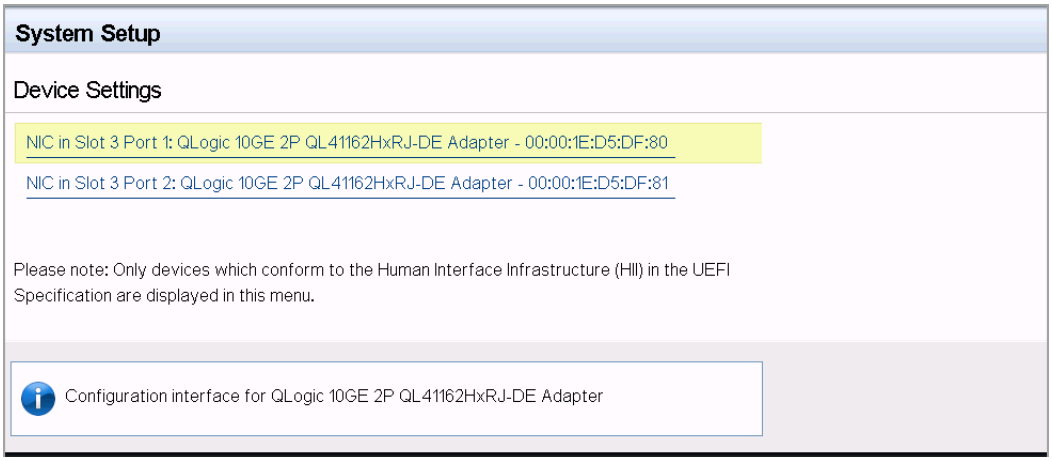


图 5-2. 系统设置：设备设置

主要配置页 （图 5-3）呈现了适配器管理选项，您可通过该选项设置分区模式。

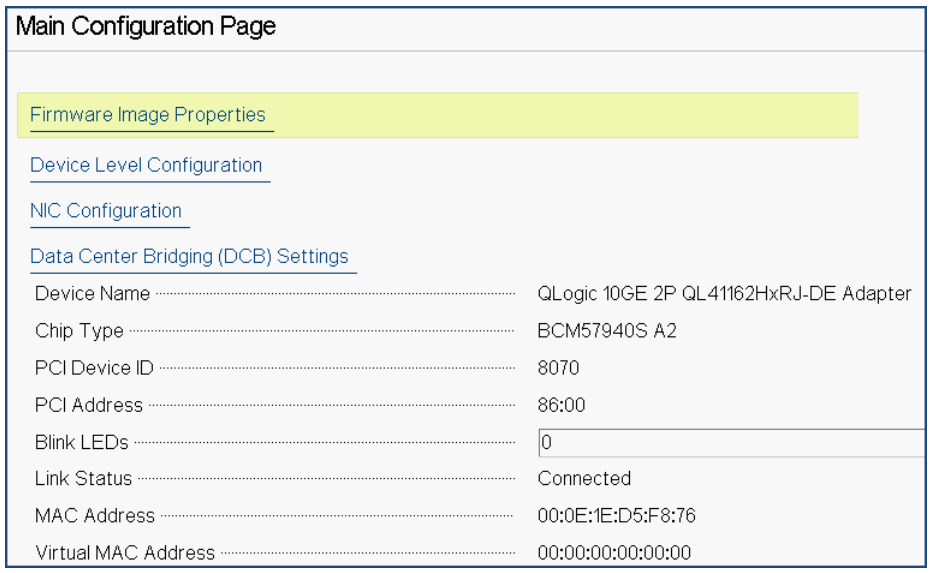


图 5-3. 主要配置页面

4. 在 **Device Level Configuration** （设备级配置）下，将 **Partitioning Mode**（分区模式）设置为 **NPAR**，以便将 **NIC Partitioning Configuration**（NIC 分区配置）选项添加到 Main Configuration Page （主要配置页面），如图 5-4 中所示。

注

NPAR 不可用于最高速度为 1G 的端口。

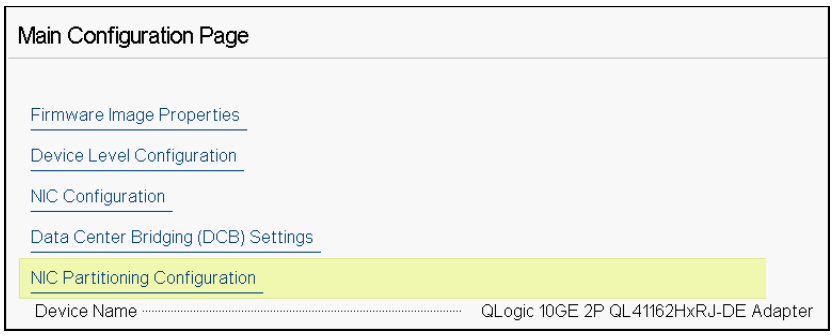


图 5-4. 主要配置页面，将分区模式设置为 NPAR

在 图 5-3 和 图 5-4 中， Main Configuration Page （主要配置页面）显示如下内容：

- **Firmware Image Properties**（固件映像属性）（请参阅第 46 页上 “ 显示固件映像属性 ”）
- **Device Level Configuration**（设备级配置）（请参阅第 47 页上 “ 配置设备级参数 ”）
- **NIC Configuration** （NIC 配置）（请参阅第 48 页上 “ 配置 NIC 参数 ”）
- **iSCSI Configuration** （iSCSI 配置）（如果通过在端口的第三个分区上以 NPAR 模式启用 iSCSI 卸载，允许 iSCSI 远程引导）（请参阅第 55 页上 “ 配置 iSCSI 引导 ”）
- **FCoE Configuration** （FCoE 配置）（如果通过在端口的第二个分区上启用 NPAR 模式下的 FCoE 卸载，允许从 SAN 的 FCoE 引导）（请参阅第 53 页上 “ 配置 FCoE 引导 ”）
- **Data Center Bridging (DCB) Settings** （数据中心桥接 （DCB）设置）（请参阅第 52 页上 “ 配置数据中心桥接 ”）
- **NIC Partitioning Configuration** （NIC 分区配置）（如果在 Device Level Configuration （设备级配置）页面上选择 **NPAR**）（请参阅第 59 页上 “ 配置分区 ”）

此外， Main Configuration Page （主要配置页面）还提供表 5-1 中所列的适配器属性。

表 5-1. 适配器属性

适配器属性	说明
Device Name （设备名称）	工厂分配的设备名称
Chip Type （芯片类型）	ASIC 版本
PCI Device ID （PCI 设备 ID）	唯一的供应商特定 PCI 设备 ID
PCI Address （PCI 地址）	采用总线 - 设备功能格式的 PCI 设备地址
Blink LEDs （闪烁 LED）	用户定义的端口 LED 闪烁次数
Link Status （链路状态）	外部链路状态
MAC Address （MAC 地址）	制造商分配的永久设备 MAC 地址
Virtual MAC Address （虚拟 MAC 地址）	用户定义的设备 MAC 地址
iSCSI MAC Address （iSCSI MAC 地址）[a]	制造商分配的永久设备 iSCSI 卸载 MAC 地址
iSCSI 虚拟 MAC 地址 [a]	用户定义的设备 iSCSI 卸载 MAC 地址

表 5-1. 适配器属性 (续)

适配器属性	说明
FCoE MAC Address （FCoE MAC 地址）[b]	制造商分配的永久设备 FCoE 卸载 MAC 地址
FCoE 虚拟 MAC 地址 [b]	用户定义的设备 FCoE 卸载 MAC 地址
FCoE WWPN [b]	制造商分配的永久设备 FCoE 卸载 WWPN （全球端口名称）
FCoE 虚拟 WWPN [b]	用户定义的设备 FCoE 卸载 WWPN
FCoE WWNN [b]	制造商分配的永久设备 FCoE 卸载 WWNN （全球节点名称）
FCoE 虚拟 WWNN [b]	用户定义的设备 FCoE 卸载 WWNN

[a] 仅当在 NIC Partitioning Configuration （NIC 分区配置）页面上启用 **iSCSI Offload** （iSCSI 卸载）后，此属性才可见。

[b] 仅当在 NIC Partitioning Configuration （NIC 分区配置）页面上启用 **FCoE Offload** （FCoE 卸载）后，此属性才可见。

显示固件映像属性

要查看固件映像的属性，请选择 Main Configuration Page （主要配置页面）上的 **Firmware Image Properties** （固件映像属性），然后按 ENTER 键。 Firmware Image Properties （固件映像属性）页面 （图 5-5）指定以下仅查看数据：

- **Family Firmware Version** （系列固件版本）是多引导映像版本，其中包含多个固件组件映像。
- **MBI Version** （MBI 版本）是在设备上处于活动状态的 Marvell FastLinQ 捆绑映像版本。
- **Controller BIOS Version** （控制器 BIOS 版本）是管理固件版本。
- **EFI Driver Version** （EFI 驱动程序版本）是可扩展固件接口 (EFI) 驱动程序版本。

- **L2B Firmware Version**（L2B 固件版本）是用于引导的 NIC 卸载固件版本。

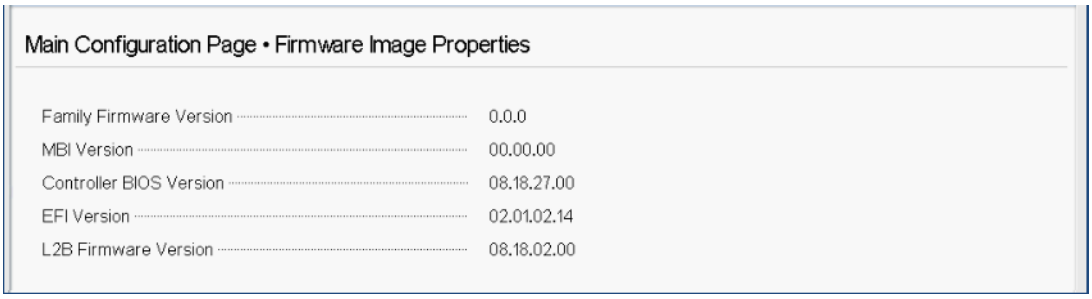


图 5-5. 固件映像属性

配置设备级参数

注

仅在 NPAR 模式下启用 iSCSI 卸载功能时，将列出 iSCSI 物理功能 (PF)。仅当在 NPAR 模式下启用 FCoE 卸载功能后，列出 FCoE PF。并非所有适配器型号都支持 iSCSI Offload （iSCSI 卸载）和 FCoE Offload （FCoE 卸载）。每个端口只能启用一个卸载，并且仅在 NPAR 模式下才能启用。

设备级配置包括以下参数：

- **Virtualization Mode** （虚拟化模式）
- **NPAREP Mode** （NPAREP 模式）

要配置设备级参数：

1. 在 Main Configuration Page （主要配置页面）上，选择 **Device Level Configuration**（设备级配置）（请参阅第 44 页上 图 5-3），然后按 ENTER 键。
2. 在 **Device Level Configuration** （设备级配置）页面上，选择设备级参数的值，如图 5-6 中所示。

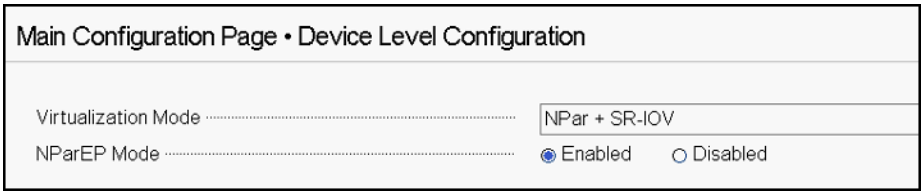


图 5-6. 设备级配置

注

QL41264HMCU-DE （部件号 5V6Y4）和 QL41264HMRJ-DE （部件号 0D1WT）适配器在设备级配置中显示支持 NPAR、 SR-IOV 和 NPAR-EP，但 1Gbps 端口 3 和 4 不支持这些功能。

3. 对于 **Virtualization Mode** （虚拟化模式），请选择以下模式之一应用于所有适配器端口：
   - ❑ **None** （无）（默认值）指定不启用虚拟化模式。
   - ❑ **NPAR** 将适配器设置为与交换机无关的 NIC 分区模式。
   - ❑ **SR-IOV** 将适配器设置为 SR-IOV 模式。
   - ❑ **NPar + SR-IOV** 通过 NPAR 模式将适配器设置为 SR-IOV。
4. **NParEP Mode** （NParEP 模式）配置每个适配器的最大分区数。如果您在步骤 2 中将 **NPAR** 或 **NPar + SR-IOV** 选为 **Virtualization Mode**（虚拟化模式），则会显示此参数。
   - ❑ **Enabled** （已启用）允许您为每个适配器配置最多 16 个分区。
   - ❑ **Disabled** （已禁用）允许您为每个适配器配置最多 8 个分区。
5. 单击 **Back** （后退）。
6. 看到提示时，单击 **Yes** （是）以保存更改。更改会在系统重设后生效。

配置 NIC 参数

NIC 配置包括设置以下参数：

- **Link Speed**
- **NIC + RDMA Mode**
- **RDMA Protocol Support**
- **Boot Mode**
- **FEC Mode**
- **Energy Efficient Ethernet**
- **Virtual LAN Mode**
- **Virtual LAN ID**

要配置 NIC 参数：

1. 在 Main Configuration Page （主要配置页面）上，选择 **NIC Configuration**（NIC 配置）（第 44 页上 图 5-3），然后单击 **Finish** （完成）。

   图 5-7 显示了 NIC Configuration （NIC 配置）页面。

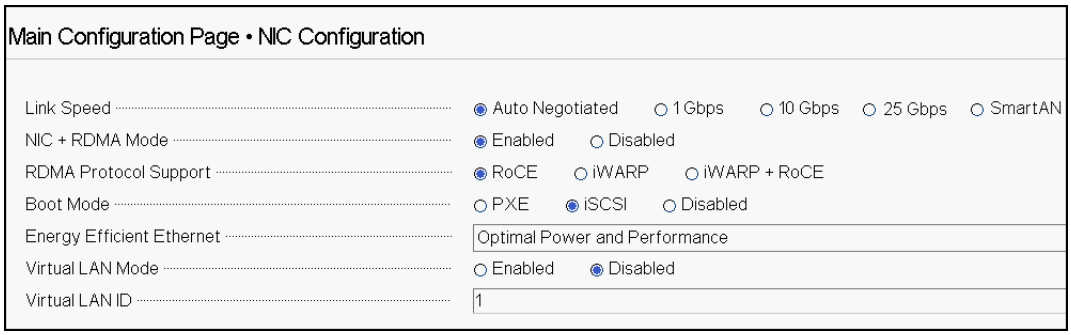


图 5-7. NIC 配置

2. 为所选端口选择以下 **Link Speed** （链路速度）选项之一。并非所有适配器都提供所有速度选择。

   - **Auto Negotiated** （自动协商）在端口上启用自动协商模式。 FEC 模式选择不可用于此速度模式。
   - **1 Gbps** 在端口上启用 1GbE 固定速度模式。此模式仅用于 1GbE 接口，不应配置用于以其他速度运行的适配器接口。 FEC 模式选择不可用于此速度模式。此模式在所有适配器上都不可用。
   - **10 Gbps** 在端口上启用 10GbE 固定速度模式。此模式在所有适配器上都不可用。
   - **25 Gbps** 在端口上启用 25GbE 固定速度模式。此模式在所有适配器上都不可用。
   - **SmartAN** （默认）在端口上启用 FastLinQ SmartAN™ 链路速度模式。 FEC 模式选择不可用于此速度模式。 **SmartAN** 设置在所有可能的链路速度和 FEC 模式之间循环，直到链路建立。该模式仅用于 25G 接口。此模式在所有适配器上都不可用。

3. 对于 **NIC + RDMA Mode** （NIC + RDMA 模式），为端口上的 RDMA 选择 **Enabled** （已启用）或 **Disabled** （已禁用）。如果处于 NPAR 模式，此设置适用于端口的所有分区。

4. 如果在 步骤 2 中将 **25 Gbps** 固定速度模式选为 **Link Speed** （链路速度），则 **FEC Mode** （FEC 模式）可见。对于 **FEC Mode** （FEC 模式），选择以下选项之一。并非所有适配器都提供所有 FEC 模式。
   - ❑ **None** （无）禁用所有 FEC 模式。
   - ❑ **Fire Code** 启用 Fire Code (BASE-R) FEC 模式。
   - ❑ **Reed Solomon** 启用 Reed Solomon FEC 模式。
   - ❑ **Auto** （自动）使端口能够以流的方式在 **None** （无）、 **Fire Code** 和 **Reed Solomon** FEC 模式之间循环 （以该链路速度），直到链路建立。
5. 如果处于 NPAR 模式，则 **RDMA Protocol Support** （RDMA 协议支持）设置适用于端口的所有分区。如果在步骤 3 中将 **NIC + RDMA Mode** （NIC + RDMA 模式）设置为 **Enabled** （启用），则会显示此设置。 **RDMA Protocol Support** （RDMA 协议支持）选项包括以下内容：
   - ❑ **RoCE** 在此端口上启用 RoCE 模式。
   - ❑ **iWARP** 在此端口上启用 iWARP 模式。
   - ❑ **iWARP + RoCE** 在此端口上启用 iWARP 和 RoCE 模式。这是默认值。此选项需要 Linux 的其他配置，如第 179 页上 “ 配置 iWARP 和 RoCE”中所述。
6. 对于 **Boot Mode** （引导模式），选择以下值之一：
   - ❑ **PXE** 启用 PXE 引导。
   - ❑ **FCoE** 通过硬件卸载路径从 SAN 启动 FCoE 引导。仅在 NPAR 模式下的第二个分区上启用 **FCoE Offload** （FCoE 卸载）后， **FCoE** 模式才可用 （请参阅第 59 页上 “ 配置分区 ”）。
   - ❑ **iSCSI** 通过硬件卸载路径启用 iSCSI 远程引导。仅在 NPAR 模式下的第三个分区上启用 **iSCSI Offload** （iSCSI 卸载）后， **iSCSI** 模式才可用 （请参阅第 59 页上 “ 配置分区 ”）。
   - ❑ **Disabled** （已禁用）防止此端口被用作远程引导源。
7. **Energy Efficient Ethernet** （节能以太网） (EEE) 参数仅在 100BASE-T 或 10GBASE-T RJ45 接口适配器上可见。从以下 EEE 选项中选择：
   - ❑ **Disabled** （已禁用）在端口上禁用 EEE。
   - ❑ **Optimal Power and Performance** （优化能耗和性能）使 EEE 在此端口上处于最佳能耗和性能模式。

- **Maximum Power Savings** （最大节能）在此端口上以最大节能模式启用 EEE。
- **Maximum Performance** （最高性能）在此端口上以最高性能模式启用 EEE。

8. 当处于 PXE 远程安装模式时，**Virtual LAN Mode**（虚拟 LAN 模式）参数适用于整个端口。 PXE 远程安装完成后，它不会持续存在。从以下 vLAN 选项中选择：
   - 对于 PXE 远程安装模式， **Enabled** （启用）在此端口上启用 vLAN 模式，
   - **Disabled** （已禁用）在此端口上禁用 vLAN 模式。
9. **Virtual LAN ID** （虚拟 LAN ID）参数指定要在此端口上用于 PXE 远程安装模式的 vLAN 标记 ID。仅当之前步骤中已启用 **Virtual LAN Mode** （虚拟 LAN 模式）时，此设置才适用。
10. 单击 **Back** （后退）。
11. 看到提示时，单击 **Yes** （是）以保存更改。更改会在系统重设后生效。

要配置使用 RDMA 的端口：

注

按照以下步骤，在 NPAR 模式端口的所有分区上启用 RDMA。

1. 将 **NIC + RDMA Mode** （NIC + RDMA 模式）设置为 **Enabled** （已启用）。
2. 单击 **Back** （后退）。
3. 看到提示时，单击 **Yes** （是）以保存更改。更改会在系统重设后生效。

要配置端口的引导模式：

1. 对于 UEFI PXE 远程安装，将 **PXE** 选为 **Boot Mode** （引导模式）。
2. 单击 **Back** （后退）。
3. 看到提示时，单击 **Yes** （是）以保存更改。更改会在系统重设后生效。

要配置端口的 PXE 远程安装以使用 vLAN：

注

PXE 远程安装完成后，此 vLAN 不会持续存在。

1. 将 **Virtual LAN Mode** （虚拟 LAN 模式）设置为 **Enabled** （已启用）。
2. 在 **Virtual LAN ID** （虚拟 LAN ID）方框中，输入使用的数字。
3. 单击 **Back** （后退）。
4. 看到提示时，单击 **Yes** （是）以保存更改。更改会在系统重设后生效。

配置数据中心桥接

数据中心桥接 (DCB) 设置包括 DCBX 协议和 RoCE 优先级。

要配置 DCB 设置：

1. 在 Main Configuration Page （主要配置页面）（第 44 页上 图 5-3）上，选择 **Data Center Bridging (DCB) Settings** （数据中心桥接 (DCB) 设置），然后单击 **Finish** （完成）。
2. 在 Data Center Bridging (DCB) Settings （数据中心桥接 (DCB) 设置）（图 5-8）页面上，选择相应的 **DCBX Protocol** （DCBX 协议）选项：
   - ❑ **Disabled** （已禁用）在端口上禁用 DCBX。
   - ❑ **CEE** 在此端口上启用旧版聚合增强型以太网 (CEE) 协议 DCBX 模式。
   - ❑ **IEEE** 在此端口上启用 IEEE DCBX 协议。
   - ❑ **Dynamic** （动态）支持动态应用 CEE 或 IEEE 协议以匹配附加的链路伙伴。

3. 在 Data Center Bridging (DCB) Settings （数据中心桥接 (DCB) 设置）页面上，将 **RoCE v1 Priority** （RoCE v1 优先级）输入为一个 **0–7** 之间的值。此设置表示用于 RoCE 流量的 DCB 流量类别优先级编号，并且应匹配用于 RoCE 流量的启用 DCB 的交换网络所用的编号。通常，0 用于默认有损流量类， 3 用于 FCoE 流量类， 4 用于通过 DCB 的无损 iSCSI-TLV 流量类。

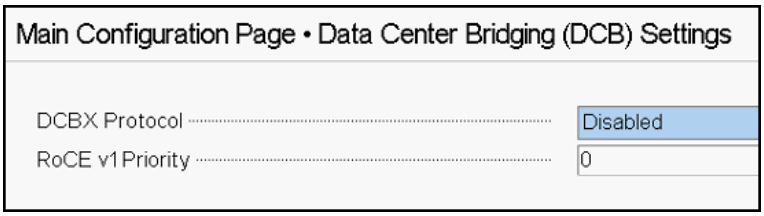


图 5-8. 系统设置：数据中心桥接 (DCB) 设置

4. 单击 **Back** （后退）。
5. 看到提示时，单击 **Yes** （是）以保存更改。更改会在系统重设后生效。

注

当启用 DCBX 时，适配器周期性地发送包含专用单播地址的链路层发现协议 (LLDP) 数据包，该专用单播地址充当源 MAC 地址。此 LLDP MAC 地址与工厂分配的适配器以太网 MAC 地址不同。如果您检查连接到适配器的交换机端口的 MAC 地址表，您将看到两个 MAC 地址：一个用于 LLDP 数据包，另一个用于适配器以太网接口。

配置 FCoE 引导

注

仅在 NPAR 模式下的第二个分区上启用 **FCoE Offload Mode** （FCoE 卸载模式）后， FCoE Boot Configuration Menu （FCoE 引导配置菜单）才可见（请参阅第 62 页上 图 5-18）。该菜单在 NPAR 模式下不可见。

要启用 FCoE 卸载模式，请参阅位于 https://www.marvell.com/documents/5aa5otcbkr0im3ynera3/ 的 *Application Note, Enabling Storage Offloads on Dell and Marvell FastLinQ 41000 Series Adapters*。

要配置 FCoE 引导配置参数：

1. 在 Main Configuration Page （主要配置页面）上，选择 **FCoE Boot Configuration** （FCoE 引导配置）菜单，然后根据需要选择以下选项：
   - ❑ **FCoE General Parameters** （FCoE 常规参数）（图 5-9）
   - ❑ **FCoE Target Configuration** （FCoE 目标配置）（图 5-10）
2. 按 ENTER 键。
3. 选择 FCoE 常规参数或 FCoE 目标配置参数的值。

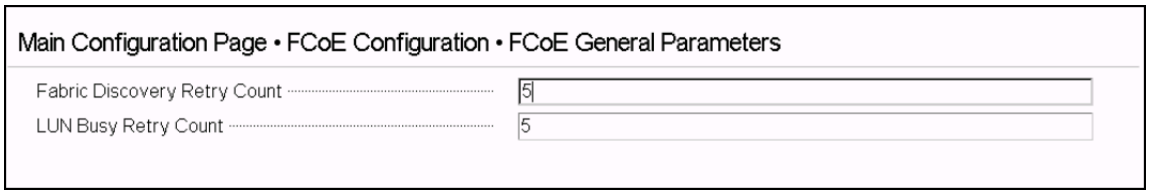


图 5-9. FCoE 常规参数

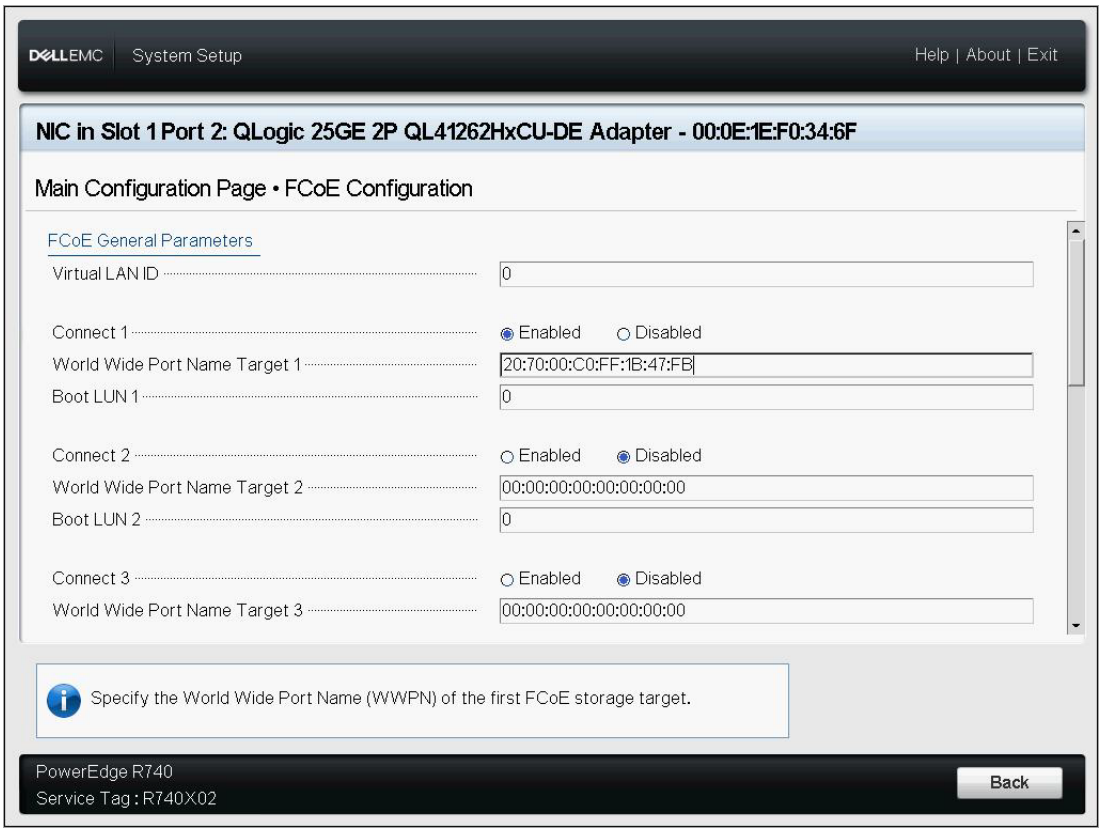


图 5-10. FCoE 目标配置

4. 单击 **Back** （后退）。
5. 看到提示时，单击 **Yes** （是）以保存更改。更改会在系统重设后生效。

配置 iSCSI 引导

注

仅在 NPAR 模式下的第三个分区上启用 **iSCSI Offload Mode** （iSCSI 卸载模式）后， iSCSI Boot Configuration Menu （FCoE 引导配置菜单）才可见（请参阅第 63 页上 图 5-19）。该菜单在 NPAR 模式下不可见。

要启用 FCoE 卸载模式，请参阅位于 https://www.marvell.com/documents/5aa5otcbkr0im3ynera3/ 的 *Application Note, Enabling Storage Offloads on Dell and Marvell FastLinQ 41000 Series Adapters*。

要配置 iSCSI 引导配置参数：

1. 在 Main Configuration Page （主要配置页面）上，选择 **iSCSI Boot Configuration** （iSCSI 引导配置）菜单，然后选择以下选项之一：
   - ❑ **iSCSI General Configuration** （iSCSI 常规配置）
   - ❑ **iSCSI Initiator Configuration** （iSCSI 启动器配置）
   - ❑ **iSCSI First Target Configuration** （iSCSI 第一目标配置）
   - ❑ **iSCSI Second Target Configuration** （iSCSI 第二目标配置）
2. 按 ENTER 键。
3. 选择相应 iSCSI 配置参数的值：
   - ❑ **iSCSI General Parameters**（iSCSI 常规参数）（第 57 页上 图 5-11）
     - ■ TCP/IP Parameters Via DHCP （通过 DHCP 获取 TCP/IP 参数）
     - ■ iSCSI Parameters Via DHCP （通过 DHCP 获取 iSCSI 参数）
     - ■ CHAP Authentication （CHAP 身份验证）
     - ■ CHAP Mutual Authentication （CHAP 相互身份验证）
     - ■ IP Version （IP 版本）
     - ■ ARP Redirect （ARP 重定向）
     - ■ DHCP Request Timeout （DHCP 请求超时）
     - ■ Target Login Timeout （目标登录超时）
     - ■ DHCP Vendor ID （DHCP 供应商 ID）

- **iSCSI Initiator Parameters** （iSCSI 启动器参数）（第 57 页上 图 5-12）
  - IPv4 Address
  - IPv4 Subnet Mask
  - IPv4 Default Gateway
  - IPv4 Primary DNS
  - IPv4 Secondary DNS
  - VLAN ID
  - iSCSI Name （iSCSI 名称）
  - CHAP ID
  - CHAP Secret （CHAP 机密）
- **iSCSI First Target Parameters** （iSCSI 第一目标参数）（第 58 页上 图 5-13）
  - Connect （连接）
  - IPv4 Address
  - TCP Port （TCP 端口）
  - Boot LUN （引导 LUN）
  - iSCSI Name （iSCSI 名称）
  - CHAP ID
  - CHAP Secret （CHAP 机密）
- **iSCSI Second Target Parameters** （iSCSI 第二目标参数）（第 58 页上 图 5-14）
  - Connect （连接）
  - IPv4 Address
  - TCP Port （TCP 端口）
  - Boot LUN （引导 LUN）
  - iSCSI Name （iSCSI 名称）
  - CHAP ID
  - CHAP Secret （CHAP 机密）

4. 单击 **Back** （后退）。
5. 看到提示时，单击 **Yes** （是）以保存更改。更改会在系统重设后生效。

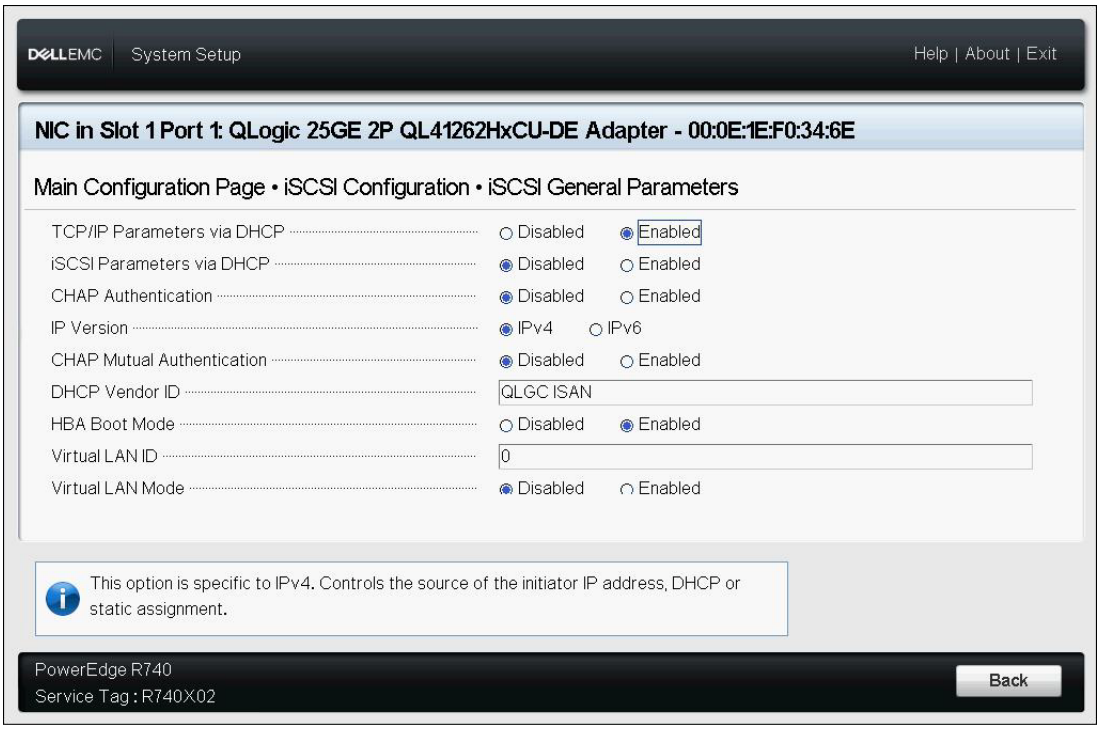


图 5-11. iSCSI 常规参数

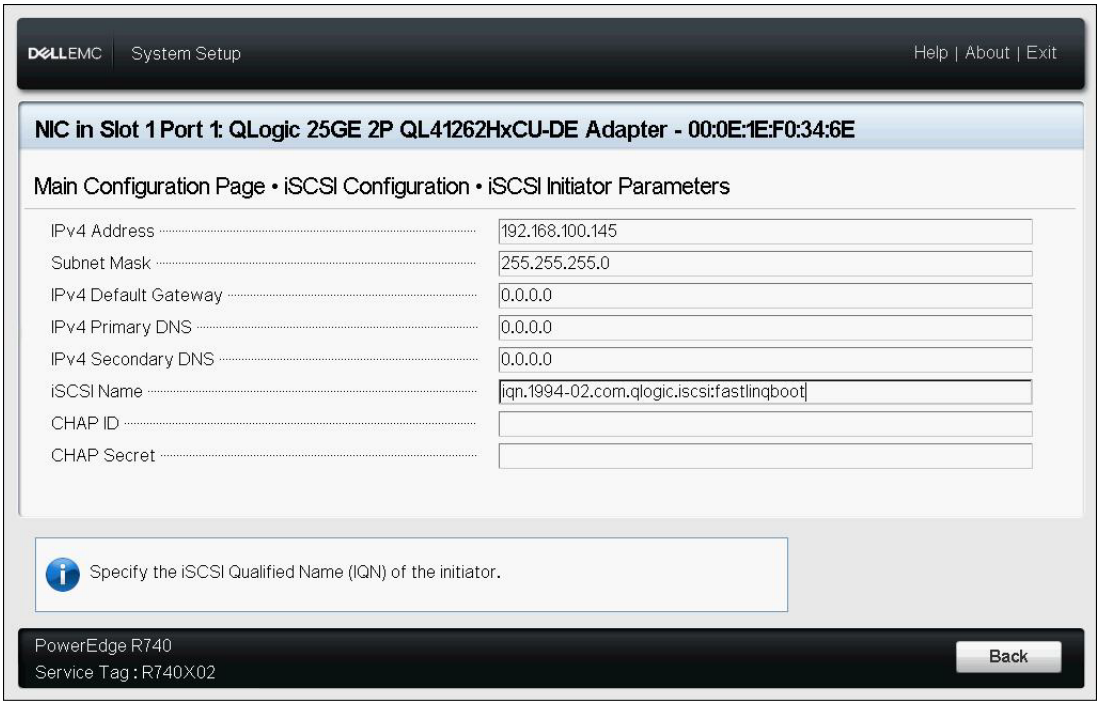


图 5-12. iSCSI 启动器配置参数

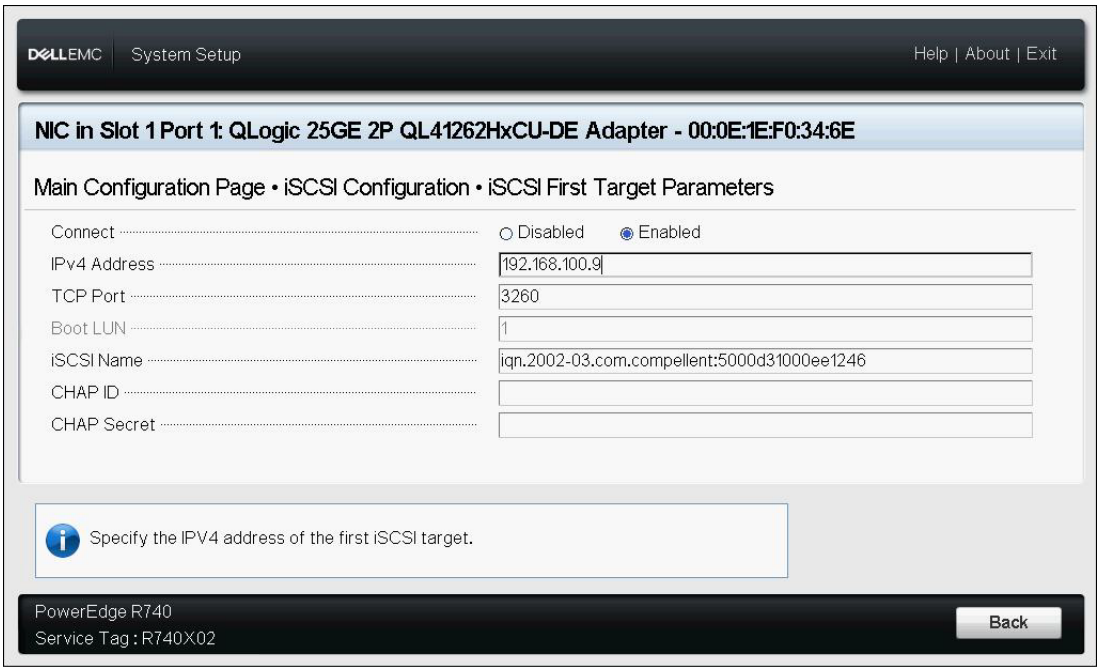


图 5-13. iSCSI 第一目标参数

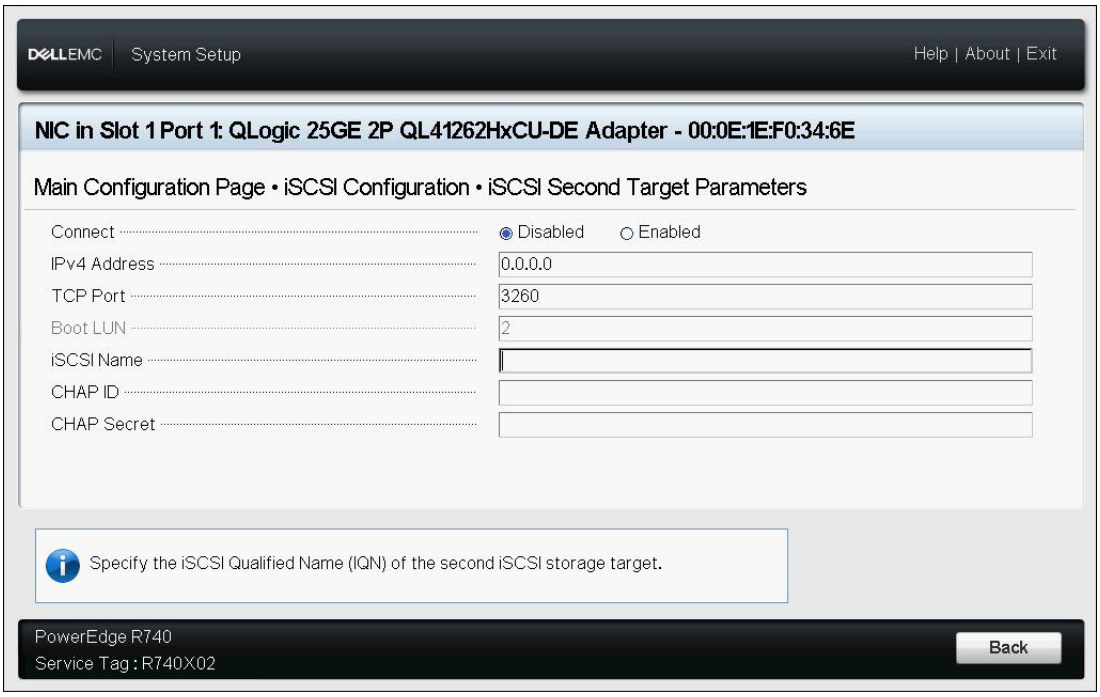


图 5-14. iSCSI 第二目标参数

配置分区

您可以为适配器上的每个分区配置带宽范围。有关 VMware ESXi 6.5 上分区配置的具体信息，请参阅对 VMware ESXi 6.5 和 ESXi 6.7 的分区。

要配置最大和最小带宽分配：

1. 在 Main Configuration Page （主要配置页面）上，选择 **NIC Partitioning Configuration** （NIC 分区配置），然后按 ENTER 键。
2. 在 Partitioning Configuration （分区配置）页面 （图 5-15）上，选择 **Global Bandwidth Allocation** （全局带宽分配）。

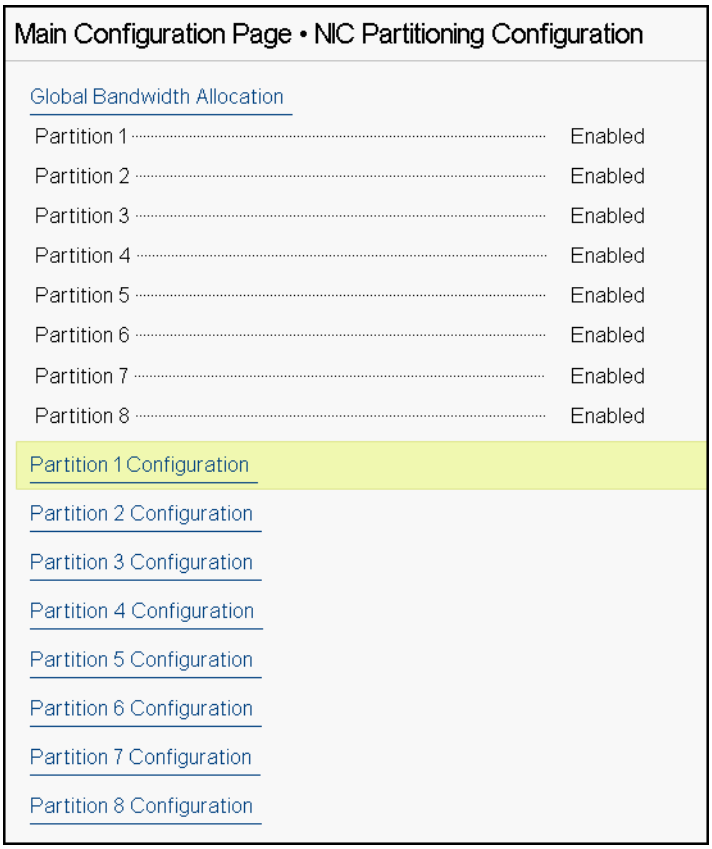


图 5-15. NIC 分区配置，全局带宽分配

3. 在 Global Bandwidth Allocation （全局带宽分配）页面 （图 5-16）上，单击您要为其分配带宽的每个分区的最小和最大 TX 带宽字段。在双端口模式下，每个端口有八个分区。

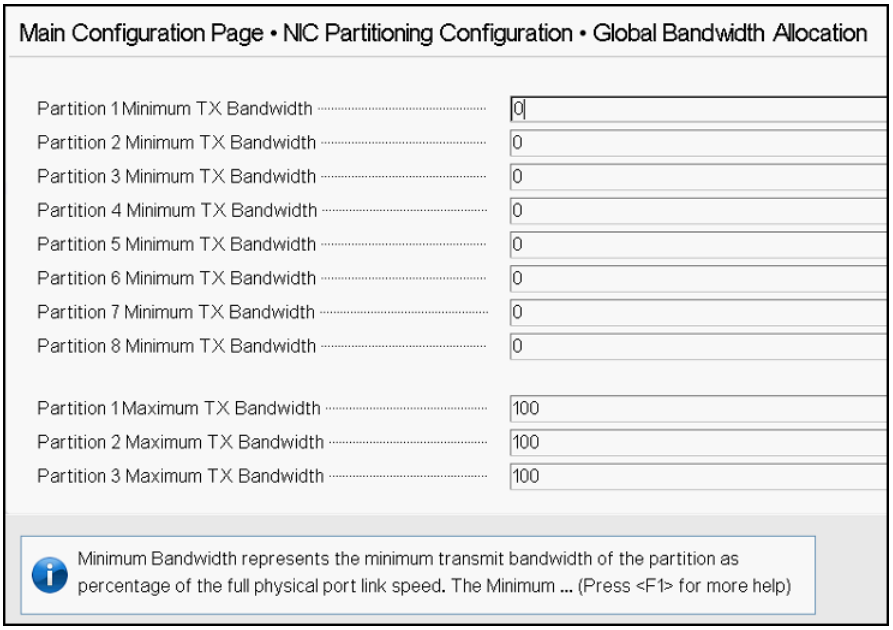


图 5-16. 全局带宽分配页面

- **Partition *n* Minimum TX Bandwidth** （分区 n 最小 TX 带宽）是所选分区的最小发送带宽，以物理端口最大链路速度的百分比形式表示。值的有效范围是 0 - 100。如果启用了 DCBX ETS 模式，则每流量类型 DCBX ETS 最小带宽值与每分区最小 TX 带宽值同时使用。单个端口上所有分区的最小 TX 带宽值的总和必须等于 100 或全为零。

  将 TX 最小带宽设置为全为零，类似于在每个活动分区上均等划分可用带宽；但是，带宽在所有主动发送的分区上动态分配。当拥塞 （来自所有分区）限制 TX 带宽时，零值 （当一个或多个其他值设置为非零值时）为该分区分配最小的百分之一。

- **Partition *n* Maximum TX Bandwidth** （分区 n 最大 TX 带宽）是所选分区的最大发送带宽，以物理端口最大链路速度的百分比形式表示。值的有效范围是 1 - 100。无论 DCBX ETS 模式设置如何，每分区最大 TX 带宽值都适用。

在每个所选字段中键入值，然后单击 **Back** （后退）。

4. 看到提示时，单击 **Yes** （是）以保存更改。更改会在系统重设后生效。

要配置分区：

1. 要检查特定分区配置，在 NIC Partitions Configuration （NIC 分区配置）页面 （第 59 页上 图 5-15）上，选择 **Partition *n* Configuration** （分区 n 配置）。如果未启用 NParEP，则每个端口仅存在四个分区。

2. 要配置第一个分区，选择 **Partition 1 Configuration** （分区 1 配置），打开 Partition 1 Configuration （分区 1 配置）页面 （图 5-17），其中显示以下参数：

   - ❑ **NIC Mode** （NIC 模式）（始终启用）
   - ❑ **PCI Device ID** （PCI 设备 ID）
   - ❑ **PCI** （总线）**地址**
   - ❑ **MAC Address** （MAC 地址）
   - ❑ **Virtual MAC Address** （虚拟 MAC 地址）

   如果未启用 NParEP，则每个端口只有四个分区可用。在非卸载功能适配器上，不会显示 **FCoE Mode** （FCoE 模式）和 **iSCSI Mode** （iSCSI 模式）选项和信息。

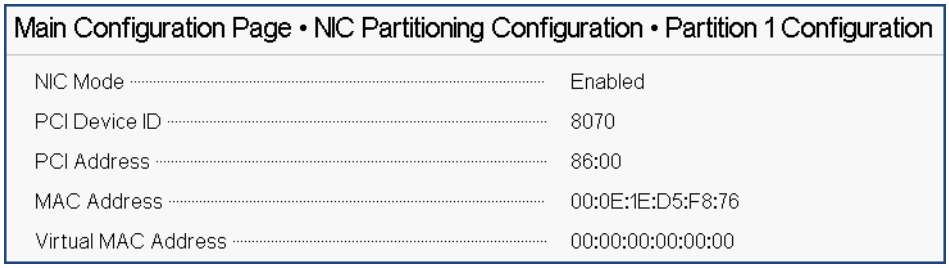


图 5-17. Partition 1 Configuration （分区 1 配置）

3. 要配置第二个分区，选择 **Partition 2 Configuration** （分区 2 配置），打开 Partition 2 Configuration （分区 2 配置）页面。如果存在 FCoE 卸载，Partition 2 Configuration （分区 2 配置）（图 5-18）会显示以下参数：

   - ❑ **NIC Mode** （NIC 模式）在分区 2 及更高版本上启用或禁用 L2 以太网 NIC 个性。要禁用任何剩余的分区，请将 **NIC Mode** （NIC 模式）设置为 **Disabled** （v 禁用）。要禁用卸载功能分区，请同时禁用 **NIC Mode** （NIC 模式）和相应的卸载模式。
   - ❑ **FCoE Mode** （FCoE 模式）在第二个分区上启用或禁用 FCoE 卸载个性。如果在第二个分区上启用此模式，则应禁用 **NIC Mode** （NIC 模式）。由于每个端口只有一个卸载可用，如果在端口的第二个分区上启用 FCoE 卸载，则无法在同一个 NPAR 模式端口的第三个分区上启用 iSCSI 卸载。并非所有适配器均支持 **FCoE Mode** （FCoE 模式）。

- **iSCSI Mode** （iSCSI 模式）在第三个分区上启用或禁用 iSCSI 卸载个性。如果在第三个分区上启用此模式，则应禁用 **NIC Mode** （NIC 模式）。由于每个端口只有一个卸载可用，如果在端口的第三个分区上启用 iSCSI 卸载，则无法在同一个 NPAR 模式端口的第二个分区上启用 FCoE 卸载。并非所有适配器均支持 **iSCSI Mode** （iSCSI 模式）。
- **FIP MAC Address** （FIP MAC 地址）[1]
- **Virtual FIP MAC Address** （虚拟 FIP MAC 地址）[1]
- **全局端口名称** [1]
- **Virtual World Wide Port Name** （虚拟全局端口名称）[1]
- **全局节点名称** [1]
- **Virtual World Wide Node Name** （虚拟全局节点名称）[1]
- **PCI Device ID （PCI 设备 ID）**
- **PCI** （总线） **Address**

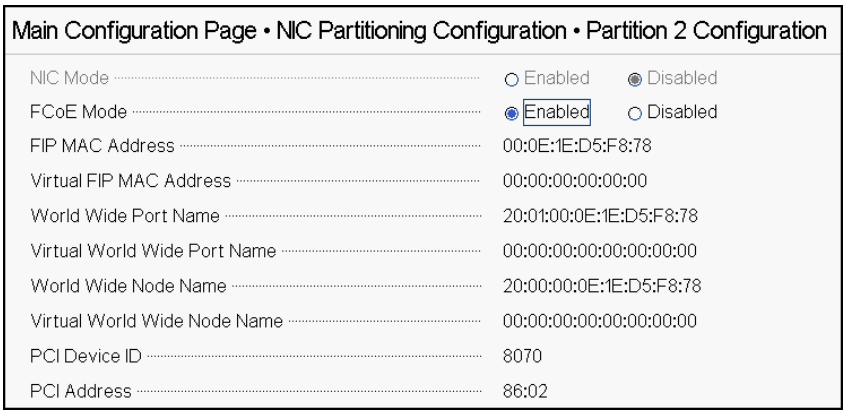


图 5-18. 分区 2 配置：FCoE 卸载

4. 要配置第三个分区，选择 **Partition 3 Configuration** （分区 3 配置），打开 Partition 3 Configuration （分区 3 配置）页面（图 5-19）如果存在 iSCSI 卸载，则 Partition 3 Configuration （分区 3 配置）会显示以下参数：

   - **NIC Mode (Disabled)** （NIC 模式 （已禁用））
   - **iSCSI Offload Mode (Enabled)** （iSCSI 卸载模式 （已启用））
   - **iSCSI Offload MAC Address** （iSCSI 卸载 MAC 地址）[2]
   - **Virtual iSCSI Offload MAC Address**（虚拟 iSCSI 卸载 MAC 地址）[2]

[1] 此参数仅存在于具有 FCoE 卸载功能的适配器的 NPAR 模式端口的第二个分区上。
[2] 此参数仅存在于具有 iSCSI 卸载功能的适配器的 NPAR 模式端口的第三个分区上。

- **PCI Device ID** （PCI 设备 ID）
- **PCI Address** （PCI 地址）

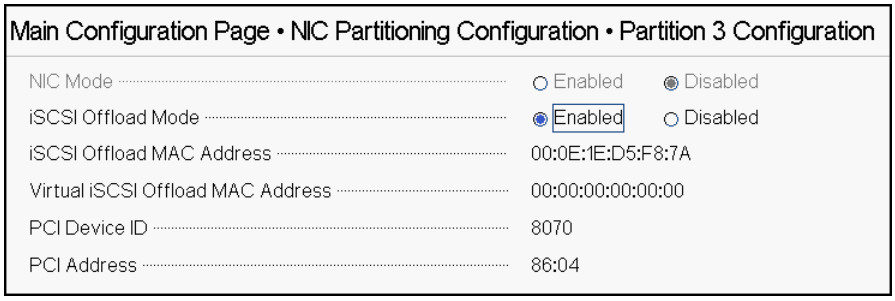


图 5-19. 分区 3 配置：iSCSI 卸载

5. 要配置剩余的以太网分区，包括以前的分区 （如果尚未启用卸载），请打开分区 2 或更大分区 （请参阅 图 5-20）页面。

   - **NIC Mode (Enabled or Disabled)** （NIC 模式 （已启用或已禁用））。禁用时，分区将被隐藏，这样如果检测到少于最大分区 （或 PCI PF）数量，OS 不会显示该分区。
   - **PCI Device ID** （PCI 设备 ID）
   - **PCI Address** （PCI 地址）
   - **MAC Address** （MAC 地址）
   - **Virtual MAC Address** （虚拟 MAC 地址）

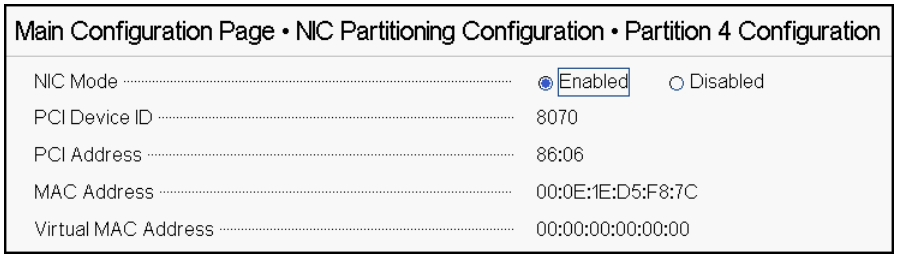


图 5-20. 分区 4 配置

对 VMware ESXi 6.5 和 ESXi 6.7 的分区

如果运行 VMware ESXi 6.5 或 ESXi 6.7 的系统上存在以下条件，则必须卸载并重新安装驱动程序：

- 将适配器配置为启用所有 NIC 分区的 NPAR。
- 适配器处于 Single Function （单功能）模式。
- 保存配置并重新引导系统。

- 如果系统上已安装驱动程序，将启用存储分区 （通过将其中一个 NIC 分区转换为存储）。
- 分区 2 变更为 FCoE。
- 保存配置并再次重新引导系统。

需要重新安装驱动程序，因为存储功能可能会保留 `vmnic`*X* 枚举而不是 `vmhba`*X*，如在系统上发出以下命令时所示：

```
# esxcfg-scsidevs -a
vmnic4  qedf              link-up   fc.2000000e1ed6fa2a:2001000e1ed6fa2a
(0000:19:00.2) QLogic Corp. QLogic FastLinQ QL41xxx Series 10/25 GbE
Controller (FCoE)
vmhba0  lsi_mr3           link-n/a  sas.51866da071fa9100
(0000:18:00.0) Avago (LSI) PERC H330 Mini
vmnic10 qedf              link-up   fc.2000000e1ef249f8:2001000e1ef249f8
(0000:d8:00.2) QLogic Corp. QLogic FastLinQ QL41xxx Series 10/25 GbE
Controller (FCoE)
vmhba1  vmw_ahci          link-n/a  sata.vmhba1
(0000:00:11.5) Intel Corporation Lewisburg SSATA Controller [AHCI mode]
vmhba2  vmw_ahci          link-n/a  sata.vmhba2
(0000:00:17.0) Intel Corporation Lewisburg SATA Controller [AHCI mode]
vmhba32 qedil            online    iscsi.vmhba32                       QLogic
FastLinQ QL41xxx Series 10/25 GbE Controller (iSCSI)
vmhba33 qedil            online    iscsi.vmhba33                       QLogic
FastLinQ QL41xxx Series 10/25 GbE Controller (iSCSI)
```

在上述命令输出中，请注意 `vmnic4` 和 `vmnic10` 实际上是存储适配器端口。为防止出现这种情况，应该在为 NPAR 模式配置适配器的同时启用存储功能。

例如，假设适配器默认处于 Single Function （单功能）模式，您应该：

1. 启用 NPAR 模式。
2. 将分区 2 更改为 FCoE。
3. 保存和重新引导。

6 从 SAN 引导配置

SAN 引导支持在引导盘位于连接至 SAN 的存储环境中部署无磁盘服务器。服务器（启动器）使用 Marvell 聚合网络适配器 (CNA) 主机总线适配器 (HBA) 通过 SAN 与存储设备 （目标）通信。

要启用 FCoE 卸载模式，请位于 https://www.marvell.com/documents/5aa5otcbkr0im3ynera3/ 的 *Application Note, Enabling Storage Offloads on Dell and Marvell FastLinQ 41000 Series Adapters*。

本章涵盖用于 iSCSI 和 FCoE 的从 SAN 引导配置：

- 从 SAN 的 iSCSI 引导
- 第 112 页上 “ 从 SAN 的 FCoE 引导 ”

从 SAN 的 iSCSI 引导

Marvell 41*xxx* 系列千兆位以太网 (GbE) 适配器支持 iSCSI 引导，从而实现无盘系统的操作系统网络引导。iSCSI 引导允许 Windows、Linux 或 VMware 操作系统通过标准 IP 网络从位于远程的 iSCSI 目标机器引导。

本节提供有关从 SAN 的 iSCSI 引导的以下配置信息：

- iSCSI 开箱即用和内建支持
- iSCSI 预引导配置
- 在 Windows 上配置从 SAN 的 iSCSI 引导
- 在 Linux 上配置从 SAN 的 iSCSI 引导
- 在 VMware 上配置 iSCSI 从 SAN 引导

iSCSI 开箱即用和内建支持

表 6-1 列出操作系统的内建和开箱即用从 SAN 的 iSCSI 引导 (BFS) 支持。

表 6-1. iSCSI 开箱即用和内建从 SAN 引导支持

OS 版本	开箱即用		内建	
	SW iSCSI BFS 支持	硬件卸载 iSCSI BFS 支持	SW iSCSI BFS 支持	硬件卸载 iSCSI BFS 支持
Windows 2012[a]	是	是	否	否
Windows 2012 R2[a]	是	是	否	否
Windows 2016[b]	是	是	是	否
Windows 2019	是	是	是	是
RHEL 7.5	是	是	是	是
RHEL 7.6	是	是	是	是
RHEL 8.0	是	是	是	是
SLES 12 SP3	是	是	是	是
SLES 15/15 SP1	是	是	是	是
vSphere ESXi 6.5 U3[c]	是	否	是	否
vSphere ESXi 6.7 U2[c]	是	否	是	否

[a] Windows Server 2012 和 2012 R2 不支持用于软硬件卸载的内建 iSCSI 驱动程序。

[b] Windows Server 2016 不支持用于硬件卸载的内建 iSCSI 驱动程序。

[c] ESXi 开箱即用和内建驱动程序不支持本地硬件卸载 iSCSI 引导。系统将执行软件引导和连接，然后将过渡到硬件卸载。

iSCSI 预引导配置

对于 Windows 和 Linux 操作系统，可以使用 **UEFI iSCSI HBA** 配置 iSCSI 引导（使用 Marvell 卸载 iSCSI 驱动程序卸载路径）。在 **Port Level Configuration**（端口级配置）下使用引导协议设置此选项。为支持 iSCSI 引导，请先在 UEFI HII 中启用 iSCSI HBA，然后相应地设置引导协议。

对于 Windows 和 Linux 操作系统，iSCSI 引导可配置为通过两种不同的方法引导：

- **iSCSI 软件**（也称为通过 Microsoft/Open-iSCSI 启动器的非卸载路径）

  遵循 iSCSI 软件安装的 Dell BIOS 指南。

- **iSCSI 硬件** （通过 Marvell FastLinQ 卸载 iSCSI 驱动程序的卸载路径）。此选项可使用 **Boot Mode** （引导模式）设置。

  iSCSI 硬件安装说明从第 69 页上 “ 启用 NPAR 和 iSCSI HBA” 开始。

对于 VMware ESXi 操作系统，只支持 iSCSI 软件方法。

本节中的 iSCSI 预引导信息包括：

- 将 BIOS 引导模式设置为 UEFI
- 启用 NPAR 和 iSCSI HBA
- 选择 iSCSI UEFI 引导协议
- 配置存储目标
- 配置 iSCSI 引导选项
- 配置 DHCP 服务器以支持 iSCSI 引导

将 BIOS 引导模式设置为 UEFI

要配置引导模式：

1. 重新启动系统。
2. 访问 System BIOS （系统 BIOS）菜单。
3. 对于 **Boot Mode** （引导模式）设置，选择 **UEFI** （请参见图 6-1）。

注

SAN 引导仅在 UEFI 环境中受支持。确保系统引导选项为 UEFI，而非旧版。

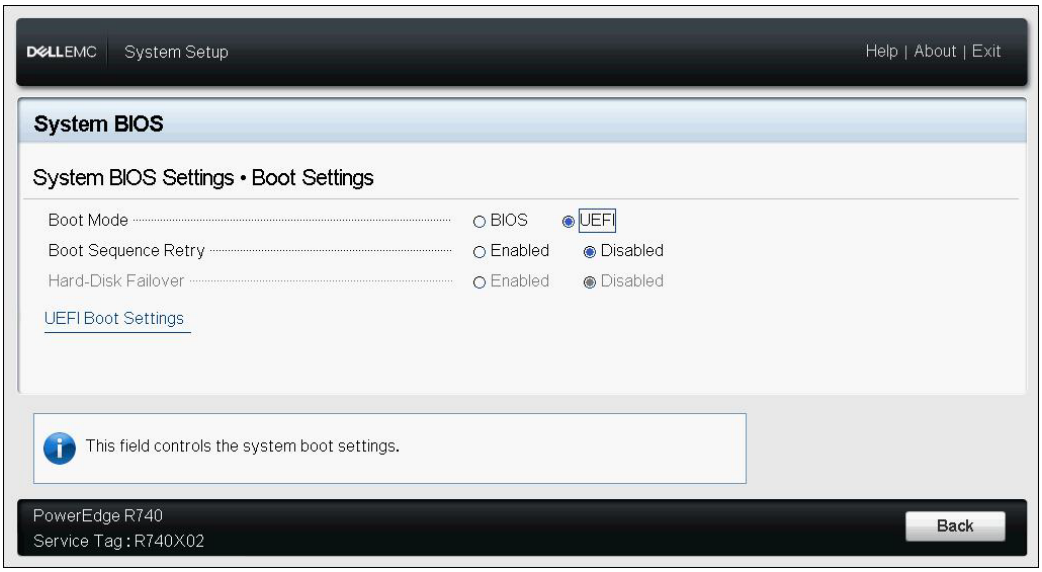


图 6-1. 系统设置：引导设置

启用 NPAR 和 iSCSI HBA

要启用 NPAR 和 iSCSI HBA：

1. 在 System Setup （系统设置）、 Device Settings （设备设置）中，选择 QLogic 设备 （图 6-2）。有关访问 PCI 设备配置菜单，请参阅 OEM 用户指南。

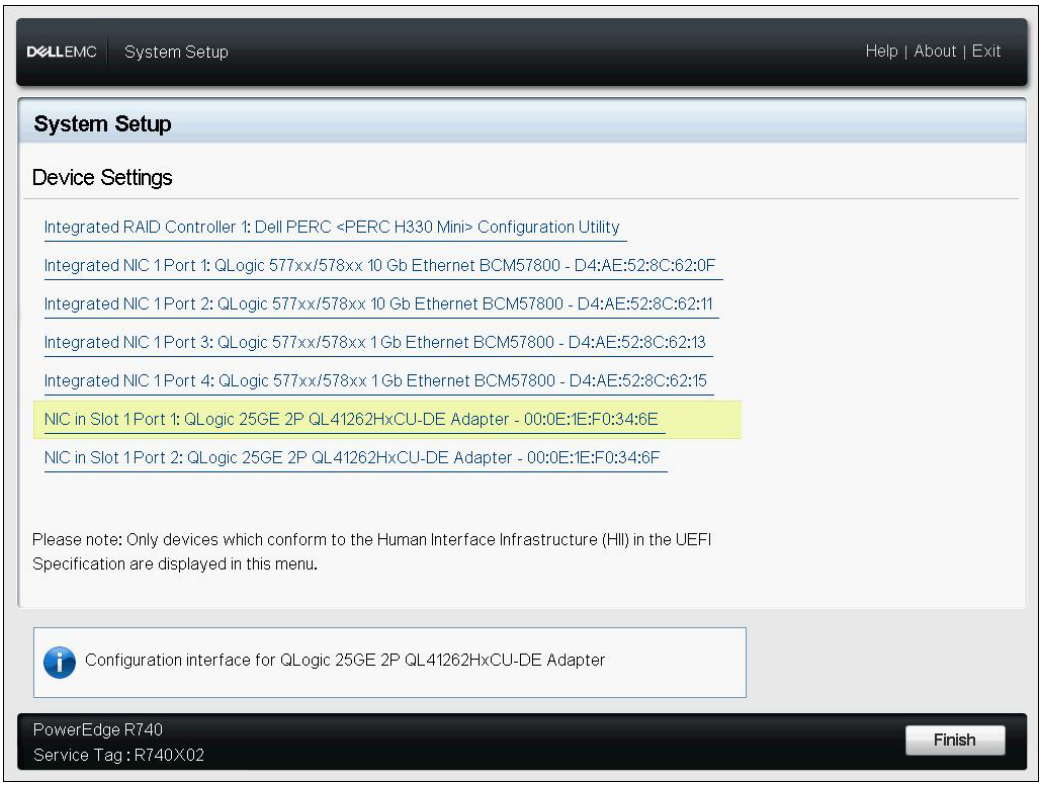


图 6-2. 系统设置：设备设置

2. 启用 NPAR。

配置存储目标

配置存储目标随目标供应商而异。有关配置存储目标的信息，请参阅供应商提供的说明文件。

要配置存储目标：

1. 基于您的存储目标选择相应的步骤：
   - ❑ 使用 SANBlaze® 或 Linux-IO (LIO™) 目标等软件创建存储目标。
   - ❑ 为 EqualLogic® 或 EMC® 等目标阵列创建虚拟盘或空间。
2. 创建一个虚拟盘。

选择 iSCSI UEFI 引导协议

在选择首选的引导模式之前，确保 **Device Level Configuration** （设备级配置）菜单设置为 **Enable NPAR** （启用 NPAR），并且 **NIC Partitioning Configuration** （NIC 分区配置）菜单设置为 **Enable iSCSI HBA** （启用 iSCSI HBA）。

Boot Mode （引导模式）选项在适配器的 **NIC Configuration** （iSCSI 配置）（图 6-3）下列出，该设置为端口特定的。有关访问 UEFI HII 下设备级配置菜单的说明，请参阅 OEM 用户手册。

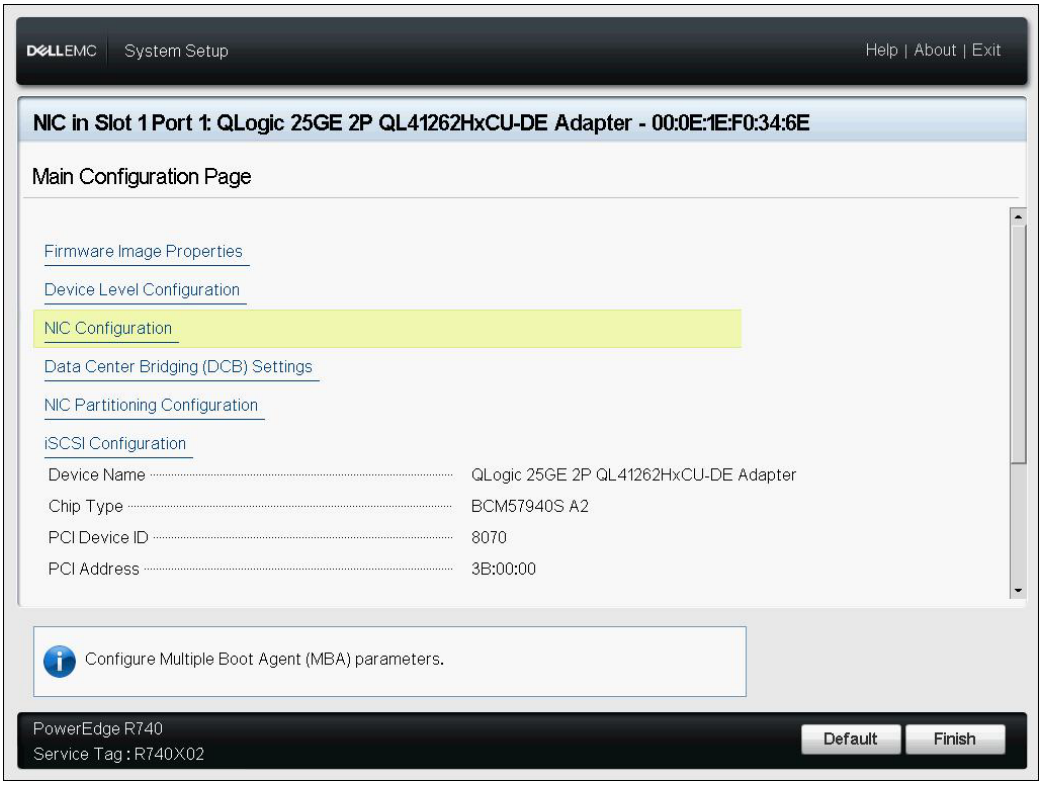


图 6-3. 系统设置：NIC 配置

注

从 SAN 的引导引导仅在 NPAR 模式下受支持，并且在 UEFI 中配置，而不是在旧版 BIOS 中配置。

1. 在 NIC 配置页面 （图 6-4）上，为 **Boot Protocol** （引导协议）选项选择 **UEFI iSCSI HBA** （需要 NPAR 模式）。

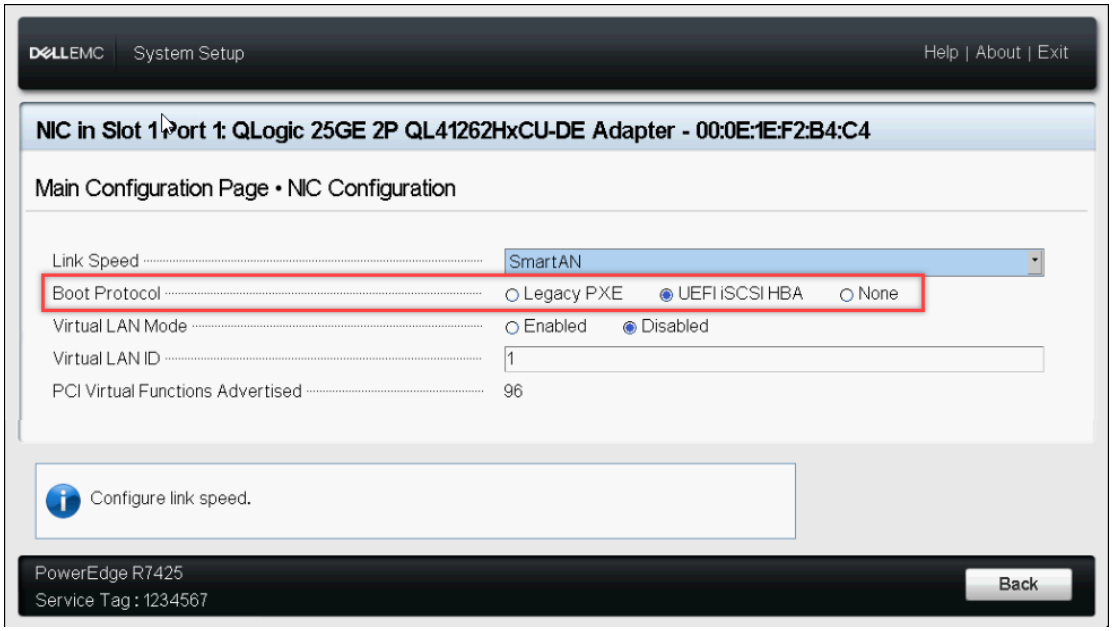


图 6-4. 系统设置：NIC 配置、引导协议

注

仅对 PXE 引导使用 **Virtual LAN Mode** （虚拟 LAN 模式）和 **Virtual LAN ID** （虚拟 LAN ID）选项。如果 UEFI iSCSI HBA 引导模式需要 vLAN，请参阅静态 iSCSI 引导配置的步骤 3。

配置 iSCSI 引导选项

iSCSI 引导配置选项包括：

- 静态 iSCSI 引导配置
- 动态 iSCSI 引导配置
- 启用 CHAP 身份验证

静态 iSCSI 引导配置

在静态配置中，您必须输入以下各项的数据：

- 启动器 IP 地址
- 启动器 IQN
- 目标参数 （在第 69 页上 “ 配置存储目标 ” 中获得）

关于配置选项的信息，请参见 第 74 页上 表 6-2。

要使用静态配置来配置 iSCSI 引导参数：

1. 在设备 HII 的 **Main Configuration Page**（主要配置页面）中，选择 **iSCSI Configuration**（iSCSI 配置）（图 6-5），然后按 ENTER 键。

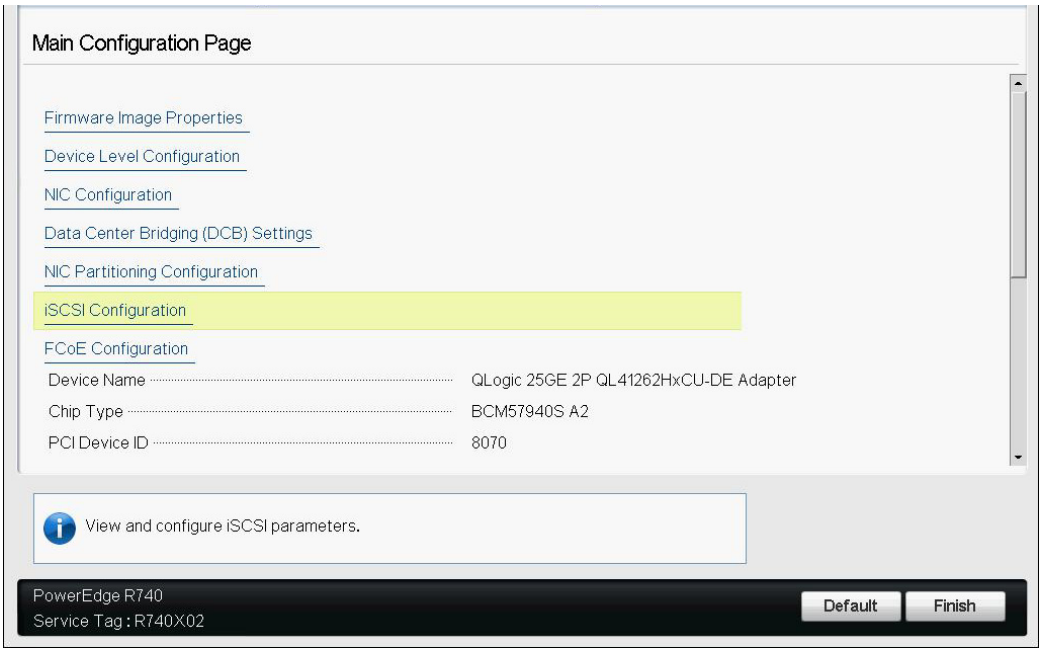


图 6-5. 系统设置：iSCSI 配置

2. 在 **iSCSI Configuration**（iSCSI 配置）页面上，选择 **iSCSI General Parameters**（iSCSI 常规参数）（图 6-6），然后按 ENTER 键。

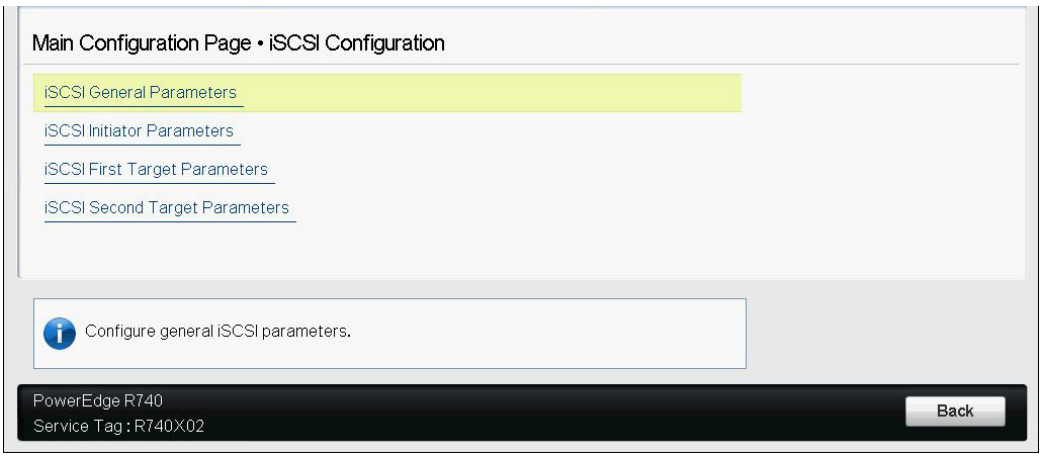


图 6-6. 系统设置：选择常规参数

3. 在 iSCSI General Parameters （iSCSI 常规参数）页面 （图 6-7）上，按向下箭头键选择参数，然后按 ENTER 键输入以下值 （第 74 页上 表 6-2 提供有这些参数的说明）：
   - **TCP/IP Parameters via DHCP** （通过 DHCP 获取 TCP/IP 参数）：**Disabled** （已禁用）
   - **iSCSI Parameters via DHCP** （通过 DHCP 获取 iSCSI 参数）：**Disabled** （已禁用）
   - **CHAP Authentication** （CHAP 身份验证）：根据需要
   - **IP Version** （IP 版本）：根据需要 （IPv4 或 IPv6）
   - **CHAP Mutual Authentication** （CHAP 身份验证）：根据需要
   - **DHCP Vendor ID** （DHCP 供应商 ID）：不适用于静态配置
   - **HBA Boot Mode** （HBA 引导模式）：根据需要
   - **Virtual LAN ID** （虚拟 LAN ID）：默认值或根据需要
   - **Virtual LAN Mode** （虚拟 LAN 模式）：根据需要

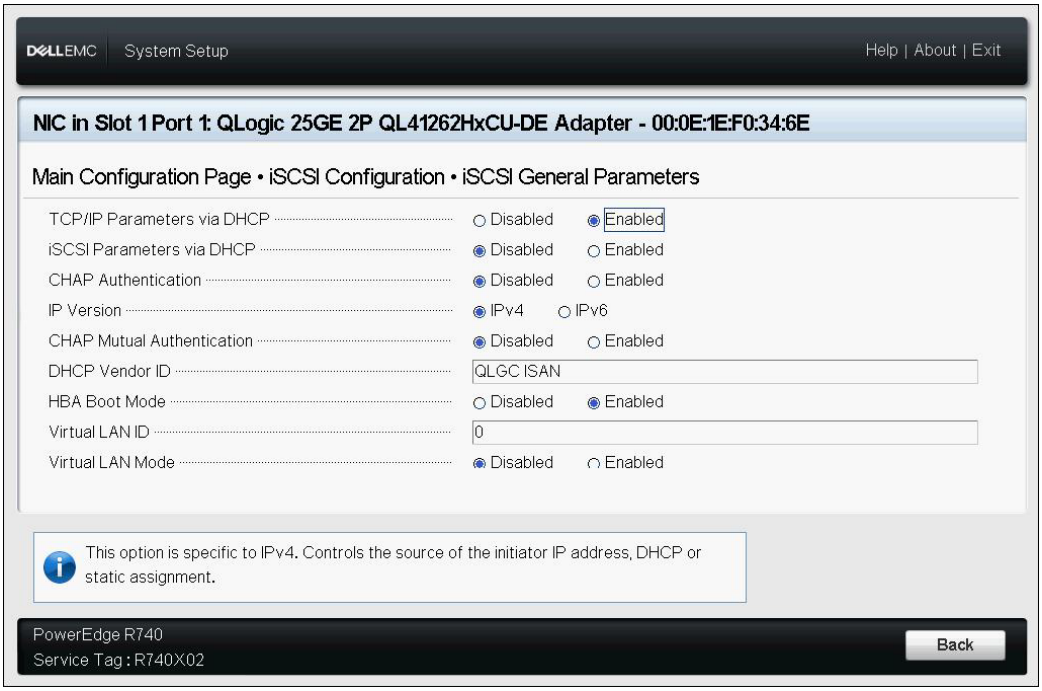


图 6-7. 系统设置：iSCSI 常规参数

表 6-2. iSCSI 常规参数

选项	说明
TCP/IP Parameters Via DHCP（通过 DHCP 获取 TCP/IP 参数）	此选项特定于 IPv4。控制 iSCSI 引导主机软件是使用 DHCP 获取 IP 地址信息（Enabled [已启用]）还是使用静态 IP 配置（Disabled [已禁用]）。
iSCSI Parameters Via DHCP（通过 DHCP 获取 iSCSI 参数）	控制 iSCSI 引导主机软件是使用 DHCP 获取其 iSCSI 目标参数（Enabled [已启用]）还是通过静态配置（Disabled [已禁用]）。静态信息在 iSCSI Initiator Parameters Configuration（iSCSI 启动器参数配置）页面上输入。
CHAP Authentication（CHAP 身份验证）	控制 iSCSI 引导主机软件在连接到 iSCSI 目标时是否使用 CHAP 身份验证。如果启用了 CHAP Authentication（CHAP 身份验证），请在 iSCSI Initiator Parameters Configuration（iSCSI 启动器参数配置）页面上配置 CHAP ID 和 CHAP Secret（CHAP 机密）。
IP Version（IP 版本）	此选项特定于 IPv6。在 IPv4 和 IPv6 之间切换。如果您从一个协议版本切换到另一个协议版本，所有 IP 设置都将丢失。
CHAP Mutual Authentication（CHAP 相互身份验证）	控制 iSCSI 引导主机软件是使用 DHCP 获取其 iSCSI 目标参数（Enabled [已启用]）还是通过静态配置（Disabled [已禁用]）。静态信息在 iSCSI Initiator Parameters Configuration（iSCSI 启动器参数配置）页面上输入。
DHCP Vendor ID（DHCP 供应商 ID）	控制 iSCSI 引导主机软件如何解释在 DHCP 期间使用的 Vendor Class ID（供应商类别 ID）字段。如果 DHCP 中的 Vendor Class ID（供应商类别 ID）字段提供的数据包匹配字段中的值，则 iSCSI 引导主机软件将查看所需 iSCSI 引导扩展中的 DHCP 选项 43 字段。如果 DHCP 被禁用，不必设置此值。
HBA Boot Mode（HBA 引导模式）	控制是启用还是禁用软件或卸载。对于卸载，此选项不可用（灰显）。有关软件（非卸载）的信息，请参阅 Dell BIOS 配置。
Virtual LAN ID（虚拟 LAN ID）	vLAN ID 范围是 1–4094。
Virtual LAN Mode（虚拟 LAN 模式）	启用或禁用 vLAN。

4. 返回到 iSCSI Configuration（iSCSI 配置）页面，然后按 ESC 键。

5. 选择 **iSCSI Initiator Parameters**（iSCSI 启动器参数）（图 6-8），然后按 ENTER 键。

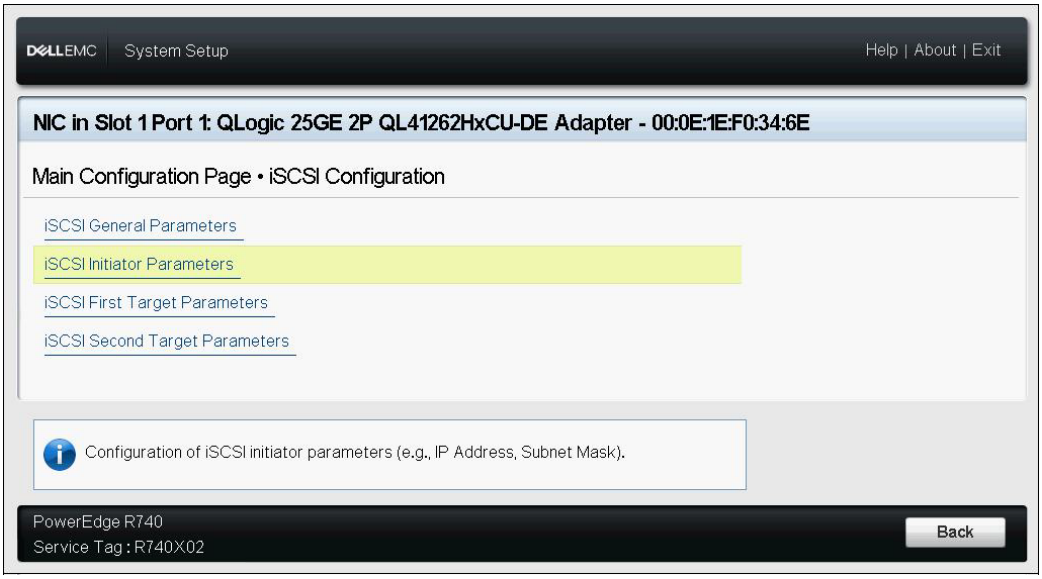


图 6-8. 系统设置：选择 iSCSI 启动器参数

6. 在 iSCSI Initiator （iSCSI 启动器）页面 （图 6-9）上，选择以下参数，然后键入各自的值：
   - ❑ **IPv4* Address**（IPv4* 地址）
   - ❑ **Subnet Mask**（子网掩码）
   - ❑ **IPv4* Default Gateway**（IPv4 默认网关）
   - ❑ **IPv4* Primary DNS**（IPv4 主 DNS）
   - ❑ **IPv4* Secondary DNS**（IPv4 辅助 DNS）
   - ❑ **iSCSI Name**（iSCSI 名称）。与客户端系统将要使用的 iSCSI 启动器名称对应。
   - ❑ **CHAP ID**
   - ❑ **CHAP Secret**（CHAP 机密）

注

要在项目前加星号 (*)，请注意以下事项：

- 标签将基于 iSCSI General Parameters （iSCSI 常规参数）页面（第 73 页上 图 6-7）上设置的 IP 版本更改为 **IPv6** 或 **IPv4** （默认值）。
- 仔细输入 IP 地址。对 IP 地址不会检查是否有重复段、错误段或网络分配错误。

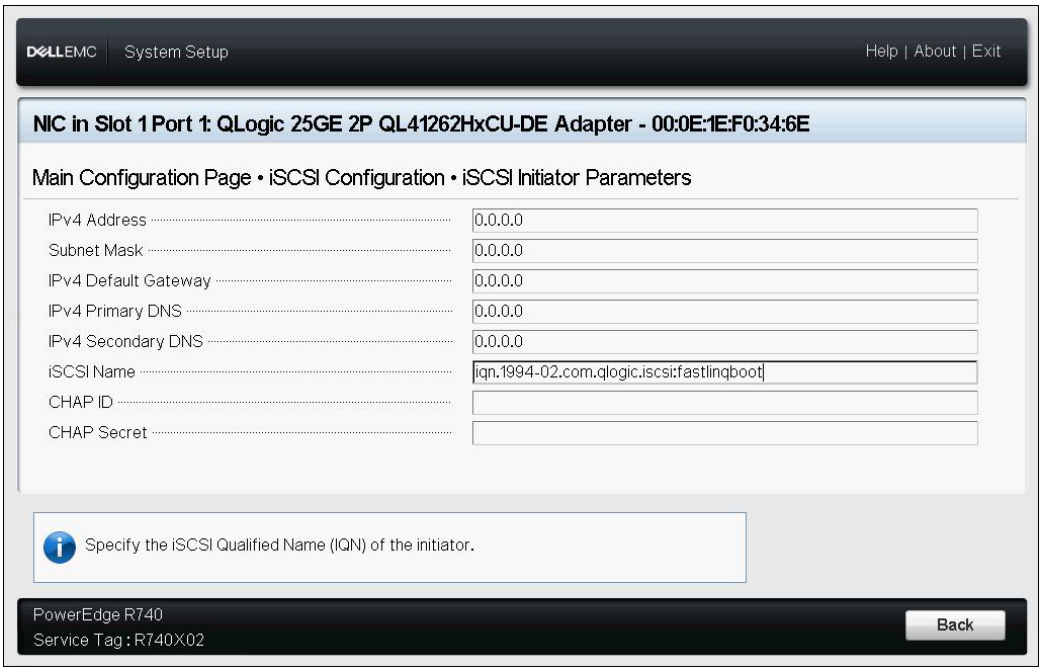


图 6-9. 系统设置：iSCSI 启动器参数

7. 返回到 iSCSI Configuration （iSCSI 配置）页面，然后按 ESC 键。

8. 选择 **iSCSI First Target Parameters** （iSCSI 第一目标参数）（图 6-10），然后按 ENTER 键。

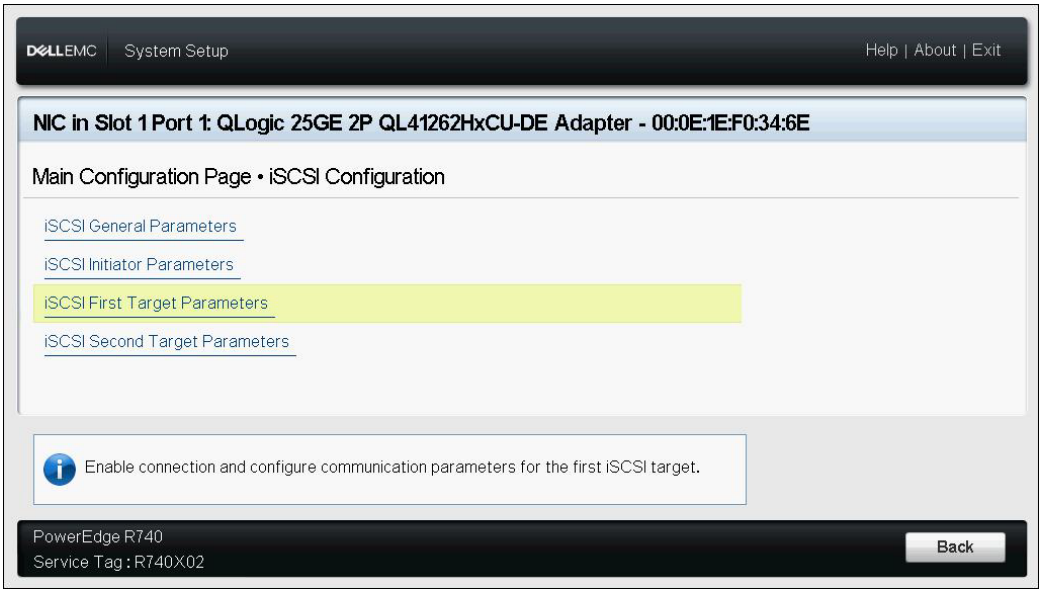


图 6-10. 系统设置：选择 iSCSI 第一目标参数

9. 在 iSCSI First Target Parameters （iSCSI 第一目标参数）页面上，将 iSCSI 目标的 **Connect** （连接）选项设置为 **Enabled** （已启用）。

10. 键入 iSCSI 目标以下参数的值，然后按 ENTER 键：

- **IPv4* Address** （IPv4* 地址）
- **TCP Port** （TCP 端口）
- **Boot LUN** （引导 LUN）
- **iSCSI Name** （iSCSI 名称）
- **CHAP ID**
- **CHAP Secret** （CHAP 机密）

注

对于上述带有星号 (*) 的参数，标签将基于 iSCSI General Parameters （iSCSI 常规参数）页面上设置的 IP 版本更改为 **IPv6** 或 **IPv4** （默认值），如图 6-11 中所示。

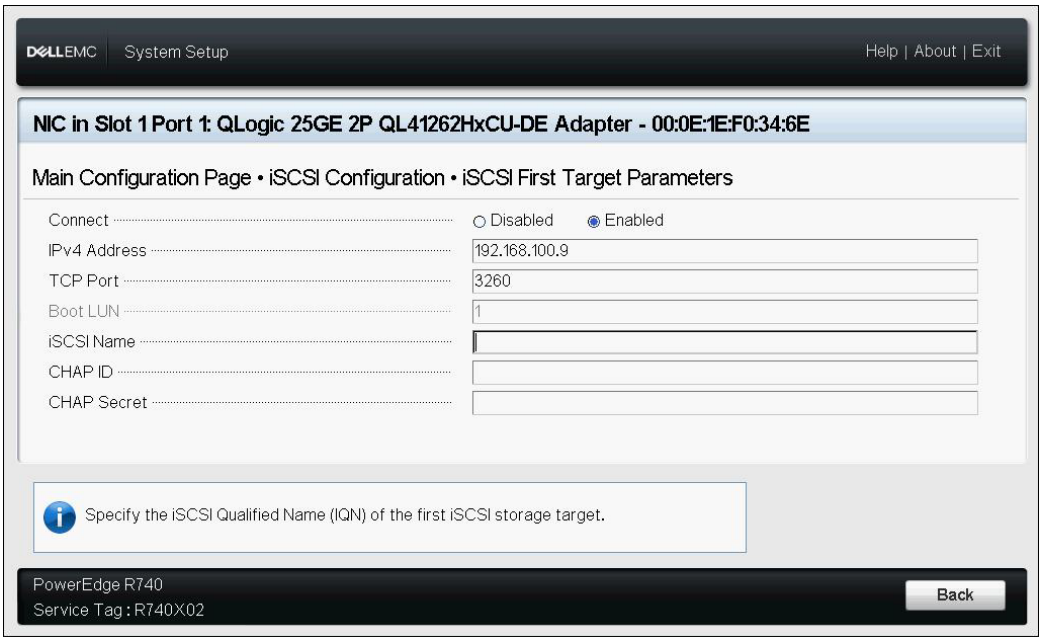


图 6-11. 系统设置：iSCSI 第一目标参数

11. 返回到 iSCSI Boot Configuration （iSCSI 引导配置）页面，然后按 ESC 键。

12. 如果要配置第二个 iSCSI 目标设备，选择 **iSCSI Second Target Parameters**（iSCSI 第二目标参数）（图 6-12），然后如同在步骤 10 中的操作一样输入参数值。如果无法连接第一个目标，则使用第二个目标。否则继续步骤 13。

图 6-12. 系统设置：iSCSI 第二目标参数

13. 按 ESC 键一次，再次按下可退出。
14. 单击 **Yes**（是）保存更改，或按照 OEM 指导操作保存设备级配置。例如，单击 **Yes**（是）确认设置更改（图 6-13）。

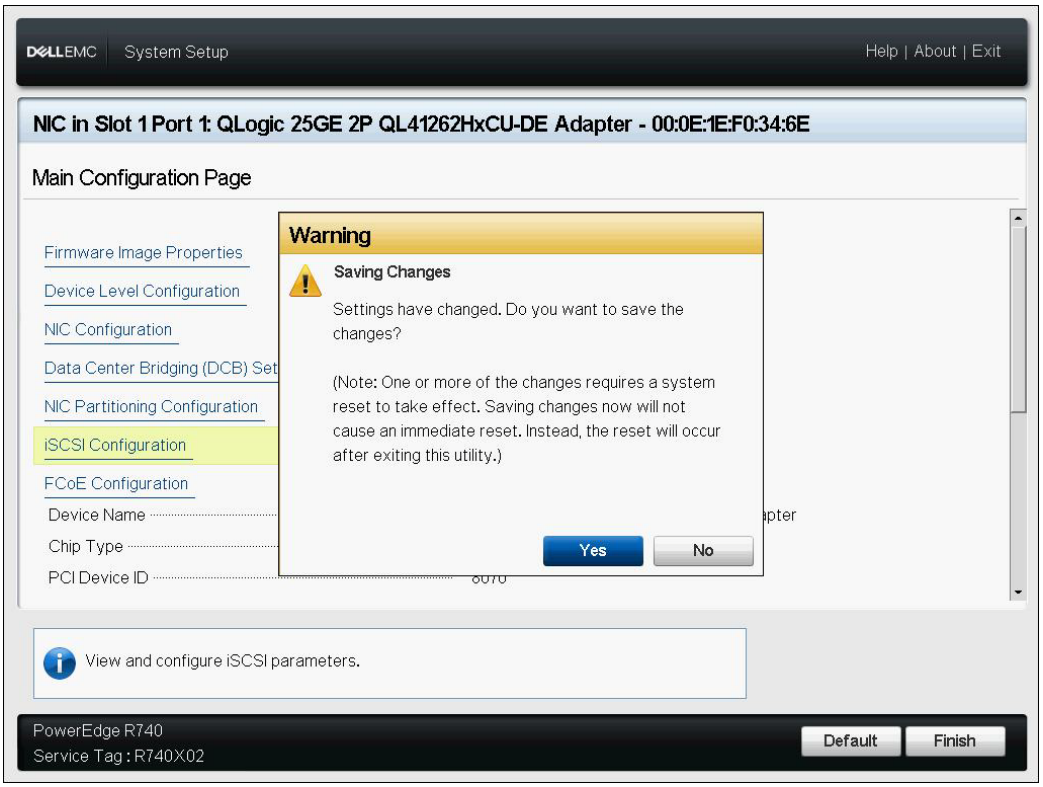


图 6-13. 系统设置：保存 iSCSI 更改

15. 进行所有更改后，重新引导系统以将更改应用到适配器正在运行的配置。

动态 iSCSI 引导配置

在动态配置中，确保系统的 IP 地址和目标 （或启动器）信息由 DHCP 服务器提供（请参阅第 83 页上 “ 配置 DHCP 服务器以支持 iSCSI 引导 ” 中的 IPv4 和 IPv6 配置）。

以下参数的任何设置都将被忽略并且无需清除 （IPv4 的启动器 iSCSI 名称、 IPv6 的 CHAP ID 和 CHAP 机密除外）：

- 启动器参数
- 第一目标参数或第二目标参数

关于配置选项的信息，请参见 第 74 页上 表 6-2。

注

使用 DHCP 服务器时， DNS 服务器条目将被 DHCP 服务器提供的值覆盖。即使本地提供的值有效并且 DHCP 服务器不提供 DNS 服务器信息，仍会发生这种覆盖。当 DHCP 服务器不提供 DNS 服务器信息时，主 DNS 服务器值和辅助 DNS 服务器值均设为 0.0.0.0。当 Windows OS 获得控制权时，Microsoft iSCSI 启动器将检索 iSCSI 启动器参数并静态配置相应的注册表。这将覆盖任何配置的参数。由于 DHCP 守护进程在 Windows 环境中作为一个用户进程运行，当堆栈在 iSCSI 引导环境中启动之前，所有 TCP/IP 参数都必须静态配置。

如果使用 DHCP 选项 17，则目标信息由 DHCP 服务器提供，且启动器 iSCSI 名称从 Initiator Parameters （启动器参数）窗口的编程值进行检索。如果未选择任何值，控制器默认名称为：

```
iqn.1995-05.com.qlogic.<11.22.33.44.55.66>.iscsiboot
```

字符串 11.22.33.44.55.66 对应于控制器的 MAC 地址。如果使用 DHCP 选项 43 （仅适用于 IPv4），则以下窗口上的任何设置都将被忽略并且无需清除：

- 启动器参数
- 第一目标参数或第二目标参数

要使用动态配置来配置 iSCSI 引导参数：

- 在 iSCSI General Parameters （iSCSI 常规参数）页面上，设置以下选项，如图 6-14 中所示：
  - **TCP/IP Parameters via DHCP** （通过 DHCP 获取 TCP/IP 参数）：Enabled （已启用）
  - **iSCSI Parameters via DHCP** （通过 DHCP 获取 iSCSI 参数）：Enabled （已启用）
  - **CHAP Authentication** （CHAP 身份验证）：根据需要
  - **IP Version** （IP 版本）：根据需要 （IPv4 或 IPv6）
  - **CHAP Mutual Authentication** （CHAP 身份验证）：根据需要
  - **DHCP Vendor ID** （DHCP 供应商 ID）：根据需要
  - **HBA Boot Mode** （HBA 引导模式）：根据需要
  - **Virtual LAN ID** （虚拟 LAN ID）：根据需要
  - **Virtual LAN Mode** （虚拟 LAN 模式）：根据需要[1]

[1] **Virtual LAN Mode** （虚拟 LAN 模式）在使用来自 DHCP 服务器的动态 （外部提供的）配置时不一定需要。

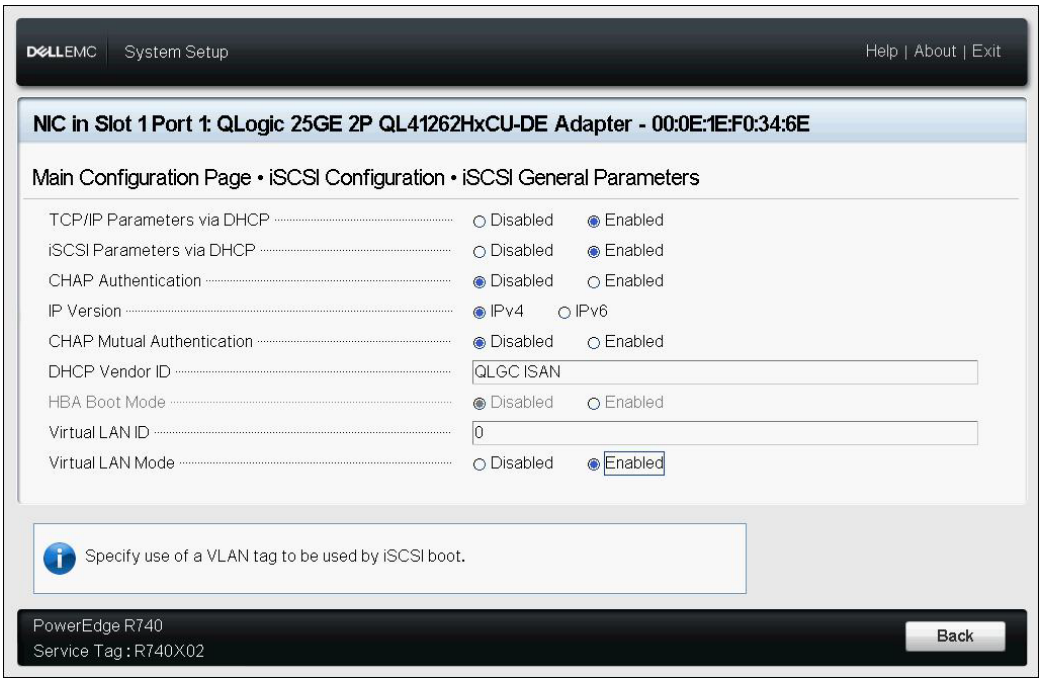


图 6-14. 系统设置：iSCSI 常规参数

启用 CHAP 身份验证

确保目标上启用 CHAP 身份验证。

要启用 CHAP 身份验证：

1. 转至 iSCSI General Parameters （iSCSI 常规参数）页面。
2. 将 **CHAP Authentication** （CHAP 身份验证）设置为 **Enabled** （已启用）。
3. 在 Initiator Parameters （启动器参数）窗口中，键入以下值：
   - **CHAP ID** （最多 255 个字符）
   - **CHAP Secret** （CHAP 机密）（如果身份验证需要；长度必须为 12 到 16 个字符）
4. 按 ESC 键返回到 iSCSI Boot Configuration （iSCSI 引导配置）页面。
5. 在 **iSCSI Boot configuration** （iSCSI 引导配置）菜单上，选择 **iSCSI First Target Parameters** （iSCSI 第一目标参数）。
6. 在 iSCSI First Target Parameters（iSCSI 第一目标参数）窗口中，键入配置 iSCSI 目标时使用的值：
   - **CHAP ID** （如果是双向 CHAP，可选填）

- ❑ **CHAP Secret** （CHAP 机密）（如果是双向 CHAP，可选填；长度必须为 12 到 16 个字符或更长）

7. 按 ESC 键返回到 iSCSI Boot Configuration （iSCSI 引导配置）菜单。
8. 按 ESC 键，然后选择确认 **Save Configuration** （保存配置）。

配置 DHCP 服务器以支持 iSCSI 引导

DHCP 服务器是一个可选组件，只有在进行动态 iSCSI 引导配置设置时才必需（请参阅第 80 页上 “ 动态 iSCSI 引导配置 ”）。

对 IPv4 和 IPv6 配置 DHCP 服务器以支持 iSCSI 引导的过程不同：

- IPv4 的 DHCP iSCSI 引导配置
- 配置 DHCP 服务器
- 为 IPv6 配置 DHCP iSCSI 引导
- 配置 vLAN 用于 iSCSI 引导

IPv4 的 DHCP iSCSI 引导配置

DHCP 包括向 DHCP 客户端提供配置信息的多个选项。对于 iSCSI 引导，Marvell FastLinQ 适配器支持以下 DHCP 配置：

- DHCP 选项 17，根路径
- DHCP 选项 43，供应商特定信息

DHCP 选项 17，根路径

选项 17 用于将 iSCSI 目标信息传递到 iSCSI 客户端。

IETC RFC 4173 中定义的根路径的格式为：

```
"iscsi:"<servername>":"<protocol>":"<port>":"<LUN>":"<targetname>"
```

表 6-3 列出 DHCP 选项 17 参数。

表 6-3. DHCP 选项 17 参数定义

参数	定义
"iscsi:"	字符串
<servername>	iSCSI 目标的 IP 地址或完全限定域名 (FQDN)
":"	分隔符
<protocol>	用于访问 iSCSI 目标的 IP 协议。由于目前仅支持 TCP，因此协议为 6。
<port>	与协议关联的端口号。 iSCSI 的标准端口号为 3260。

表 6-3. DHCP 选项 17 参数定义 (续)

参数	定义
<LUN>	要在 iSCSI 目标上使用的逻辑单元号。 LUN 的值必须以十六进制格式表示。 ID 为 64 的 LUN 在 DHCP 服务器上的选项 17 参数内必须配置为 40。
<targetname>	IQN 或 EUI 格式的目标名称。有关 IQN 和 EUI 格式的详细信息，请参阅 RFC 3720。 IQN 名称示例： iqn.1995-05.com.QLogic:iscsi-target。

DHCP 选项 43，供应商特定信息

DHCP 选项 43 （供应商特定信息）为 iSCSI 客户端提供比 DHCP 选项 17 更多的配置选项。在此配置中，还提供三个额外的子选项，将可用于引导的启动器 IQN 以及两个 iSCSI 目标 IQN 分配给 iSCSI 引导客户端。 iSCSI 目标 IQN 的格式与 DHCP 选项 17 相同，而 iSCSI 启动器 IQN 仅仅是启动器的 IQN。

注

DHCP 选项 43 仅在 IPv4 中受支持。

表 6-4 列出 DHCP 选项 43 子选项。

表 6-4. DHCP 选项 43 子选项定义

子选项	定义
201	标准根路径格式中的第一 iSCSI 目标信息： "iscsi:"<servername>":"<protocol>":"<port>":"<LUN>": "<targetname>"
202	标准根路径格式中的第二 iSCSI 目标信息： "iscsi:”<servername>":"<protocol>":"<port>":"<LUN>": "<targetname>"
203	iSCSI 启动器 IQN

使用 DHCP 选项 43 需要比 DHCP 选项 17 更多的配置，但它提供更丰富的环境和更多的配置选项。您应该在执行动态 iSCSI 引导配置时使用 DHCP 选项 43。

配置 DHCP 服务器

配置 DHCP 服务器以支持选项 16、 17 或 43。

注

DHCPv6 选项 16 和选项 17 的格式在 RFC 3315 中全面定义。

如果您使用选项 43，还必须配置选项 60。选项 60 的值必须匹配 DHCP Vendor ID （DHCP 供应商 ID）值 QLGC ISAN，如 iSCSI Boot Configuration （iSCSI 引导配置）页面的 **iSCSI General Parameters** （iSCSI 常规参数）中所示。

为 IPv6 配置 DHCP iSCSI 引导

DHCPv6 服务器可提供多个选项，包括无状态或有状态 IP 配置，以及 DHCPv6 客户端的信息。对于 iSCSI 引导， Marvell FastLinQ 适配器支持以下 DHCP 配置：

- DHCPv6 选项 16，供应商类别选项
- DHCPv6 选项 17，供应商特定信息

注

DHCPv6 标准根路径选项尚不可用。 Marvell 建议对动态 iSCSI 引导 IPv6 支持使用选项 16 或选项 17。

DHCPv6 选项 16，供应商类别选项

DHCPv6 选项 16 （供应商类别选项）必须存在且必须包含匹配您配置的 DHCP Vendor ID（DHCP 供应商 ID）参数的字符串。DHCP Vendor ID（DHCP 供应商 ID）值为 QLGC ISAN，如 iSCSI Boot Configuration （iSCSI 引导配置）菜单的 **General Parameters** （常规参数）中所示。

选项 16 的内容应为 `<2-byte length> <DHCP Vendor ID>`。

DHCPv6 选项 17，供应商特定信息

DHCPv6 选项 17 （供应商特定信息）为 iSCSI 客户端提供更多的配置选项。在此配置中，还提供三个额外的子选项，将可用于引导的启动器 IQN 以及两个 iSCSI 目标 IQN 分配给 iSCSI 引导客户端。

表 6-5 列出 DHCP 选项 17 子选项。

表 6-5. DHCP 选项 17 子选项定义

子选项	定义
201	标准根路径格式中的第一 iSCSI 目标信息： `"iscsi:"[<servername>]":"<protocol>":"<port>":"<LUN> ":"<targetname>"`
202	标准根路径格式中的第二 iSCSI 目标信息： `"iscsi:"[<servername>]":"<protocol>":"<port>":"<LUN> ":"<targetname>"`
203	iSCSI 启动器 IQN

方括号 [ ] 是 IPv6 地址所必需。

选项 17 的格式应为：

```
<2-byte Option Number 201|202|203> <2-byte length> <data>
```

配置 vLAN 用于 iSCSI 引导

网络上的 iSCSI 流量可以隔离在第 2 层 vLAN 中，以与常规流量隔离开来。如果是这种情况，请让适配器上的 iSCSI 接口成为该 vLAN 的成员。

要配置 vLAN 用于 iSCSI 引导：

1. 转至端口的 **iSCSI Configuration** （iSCSI 配置）页面。
2. 选择 **iSCSI General Parameters** （iSCSI 常规参数）。

3. 选择 **VLAN ID** 以输入和设置 vLAN 值，如图 6-15 中所示。

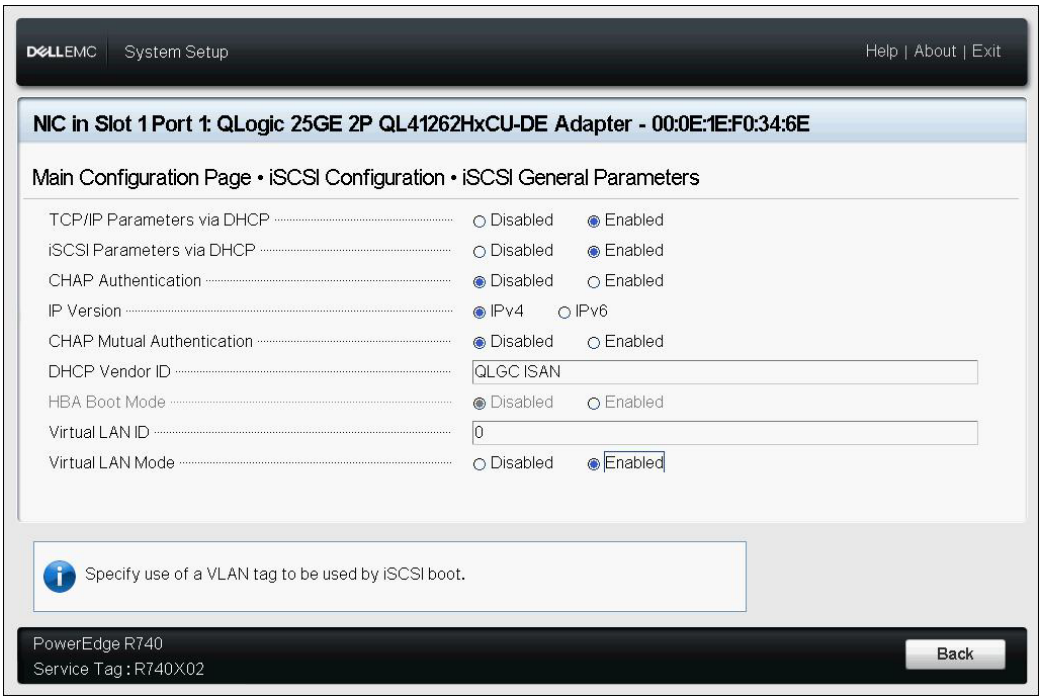


图 6-15. 系统设置：iSCSI 常规参数， VLAN ID

在 Windows 上配置从 SAN 的 iSCSI 引导

适配器支持 iSCSI 引导，从而实现无盘系统的操作系统网络引导。iSCSI 引导允许 Windows 操作系统通过标准 IP 网络从位于远程的 iSCSI 目标计算机引导。您可以打开 **NIC Configuration** （NIC 配置）菜单并将 **Boot Protocol** （引导协议）设置为 **UEFI iSCSI**，设置 L4 iSCSI 选项 （使用 Marvell 卸载 iSCSI 驱动程序的卸载路径）。

对于 Windows，从 SAN 的 iSCSI 引导信息包括以下内容：

- 准备工作
- 选择首选的 iSCSI 引导模式
- 配置 iSCSI 常规参数
- 配置 iSCSI 启动器
- 配置 iSCSI 目标
- 检测 iSCSI LUN 并注入 Marvell 驱动程序

准备工作

在 Windows 机器上开始配置从 SAN 的 iSCSI 引导之前，请注意以下事项：

- 仅使用 **NParEP 模式**的 NPAR 支持 iSCSI 引导。在配置 iSCSI 引导之前：
  1. 访问 Device Level Configuration （设备级配置）页面。
  2. 将 **Virtualization Mode** （虚拟化模式）设置为 **Npar** (NPAR)。
  3. 将 **NParEP Mode** （NParEP 模式）设置为 **Enabled** （已启用）。
- 服务器引导模式必须为 UEFI。
- 旧版 BIOS 中不支持 41*xxx* 系列适配器上的 iSCSI 引导。
- Marvell 建议禁用集成式 RAID 控制器。

选择首选的 iSCSI 引导模式

要选择 Windows 上的 iSCSI 引导模式：

1. 在所选分区的 NIC Partitioning Configuration （NIC 分区配置）页面上，将 **iSCSI Offload Mode** （iSCSI 卸载模式）设置为 **Enabled** （已启用）。
2. 在 NIC Configuration （NIC 配置）页面上，设置 **Boot Protocol** （引导协议）选项为 **UEFI iSCSI HBA**。

配置 iSCSI 常规参数

配置 Marvell iSCSI 引导软件以实现静态或动态配置。有关 General Parameters （常规参数）窗口提供的配置选项，请参阅第 74 页上 表 6-2，其中列出了适用于 IPv4 和 IPv6 的参数。

要在 Windows 上设置 iSCSI 常规参数：

1. 从主要配置页面选择 **iSCSI Configuration** （iSCSI 配置），然后选择 **iSCSI General Parameters** （iSCSI 常规参数）。
2. 在 iSCSI General Parameters （iSCSI 常规参数）页面 （参见第 73 页上 图 6-7）上，按向下箭头键选择参数，然后按 ENTER 键输入以下值 （第 74 页上 表 6-2 提供有这些参数的说明）：
   - **TCP/IP Parameters via DHCP** （通过 DHCP 获取 TCP/IP 参数）：**Disabled** （已禁用）（对于静态 iSCSI 引导）或 **Enabled** （已启用）（对于动态 iSCSI 引导）
   - **iSCSI Parameters via DHCP** （通过 DHCP 获取 iSCSI 参数）：**Disabled** （已禁用）
   - **CHAP Authentication** （CHAP 身份验证）：根据需要
   - **IP Version** （IP 版本）：根据需要 （IPv4 或 IPv6）

- ❑ **Virtual LAN ID** （虚拟 LAN ID）：（可选）您可以将网络上的 iSCSI 流量隔离在第 2 层 vLAN 中，以与常规流量隔离开来。要隔离流量，请设置此值，让适配器上的 iSCSI 接口成为第 2 层 vLAN 的成员。

配置 iSCSI 启动器

要在 Windows 上设置 iSCSI 启动器参数：

1. 从主要配置页面选择 **iSCSI Configuration** （iSCSI 配置），然后选择 **iSCSI Initiator Parameters** （iSCSI 启动器参数）。
2. 在 iSCSI Initiator Parameters （iSCSI 启动器参数）页面 （参见第 76 页上 图 6-9）上，选择以下参数，然后键入各自的值：
   - ❑ **IPv4* Address** （IPv4* 地址）
   - ❑ **Subnet Mask** （子网掩码）
   - ❑ **IPv4* Default Gateway** （IPv4 默认网关）
   - ❑ **IPv4* Primary DNS** （IPv4 主 DNS）
   - ❑ **IPv4* Secondary DNS** （IPv4 辅助 DNS）
   - ❑ **Virtual LAN ID** （虚拟 LAN ID）：（可选）您可以将网络上的 iSCSI 流量隔离在第 2 层 vLAN 中，以与常规流量隔离开来。要隔离流量，请设置此值，让适配器上的 iSCSI 接口成为第 2 层 vLAN 的成员。
   - ❑ **iSCSI Name** （iSCSI 名称）。与客户端系统将要使用的 iSCSI 启动器名称对应。
   - ❑ **CHAP ID**
   - ❑ **CHAP Secret** （CHAP 机密）

   注

   要在项目前加星号 (*)，请注意以下事项：
   - ■ 标签将基于 iSCSI General Parameters （iSCSI 常规参数）页面（参见第 73 页上 图 6-7）上设置的 IP 版本更改为 **IPv6** 或 **IPv4**（默认值）。
   - ■ 仔细输入 IP 地址。对 IP 地址不会检查是否有重复段、错误段或网络分配错误。
3. 选择 **iSCSI First Target Parameters** （iSCSI 第一目标参数）（第 77 页上 图 6-10），然后按 ENTER 键。

配置 iSCSI 目标

您可以设置 iSCSI 第一目标、第二目标或同时设置两者。

要在 Windows 上设置 iSCSI 目标参数：

1. 从主要配置页面选择 **iSCSI Configuration**（iSCSI 配置），然后选择 **iSCSI First Target Parameters**（iSCSI 第一目标参数）。
2. 在 iSCSI First Target Parameters（iSCSI 第一目标参数）页面上，将 iSCSI 目标的 **Connect**（连接）选项设置为 **Enabled**（已启用）。
3. 键入 iSCSI 目标以下参数的值，然后按 ENTER 键：
   - ❑ **IPv4* Address**（IPv4* 地址）
   - ❑ **TCP Port**（TCP 端口）
   - ❑ **Boot LUN**（引导 LUN）
   - ❑ **iSCSI Name**（iSCSI 名称）
   - ❑ **CHAP ID**
   - ❑ **CHAP Secret**（CHAP 机密）

   注

   对于上述带有星号 (*) 的参数，标签将基于 iSCSI General Parameters（iSCSI 常规参数）页面上设置的 IP 版本更改为 **IPv6** 或 **IPv4**（默认值），如第 73 页上 图 6-7 中所示。
4. 如果要配置第二个 iSCSI 目标设备，选择 **iSCSI Second Target Parameters**（iSCSI 第二目标参数）（第 79 页上 图 6-12），然后如同在步骤 3 中的操作一样输入参数值。如果无法连接第一个目标，则使用第二个目标。否则继续步骤 5。
5. 在 Warning（警告）对话框中，单击 **Yes**（是）保存更改，或按照 OEM 指导操作保存设备级配置。

检测 iSCSI LUN 并注入 Marvell 驱动程序

1. 重新引导系统，访问 HII，并确定是否检测到 iSCSI LUN。发出以下 UEFI Shell（第 2 版）脚本命令：

```
map -r -b
```

上述命令的输出如图 6-16 所示，表示在预引导级别已成功检测到 iSCSI LUN。

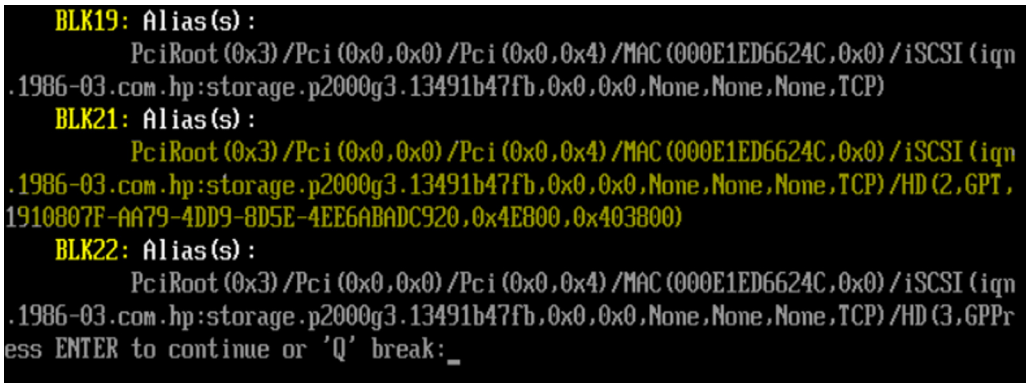


图 6-16. 使用 UEFI Shell （第 2 版）检测 iSCSI LUN

2. 在新检测到的 iSCSI LUN 中，选择安装来源，如使用 WDS 服务器、集成式 Dell 远程访问控制器 (iDRAC) 的安装 `.ISO` 或使用 CD/DVD。
3. 在 Windows 设置窗口 （图 6-17）中，选择要在其中安装驱动程序的驱动器名称。

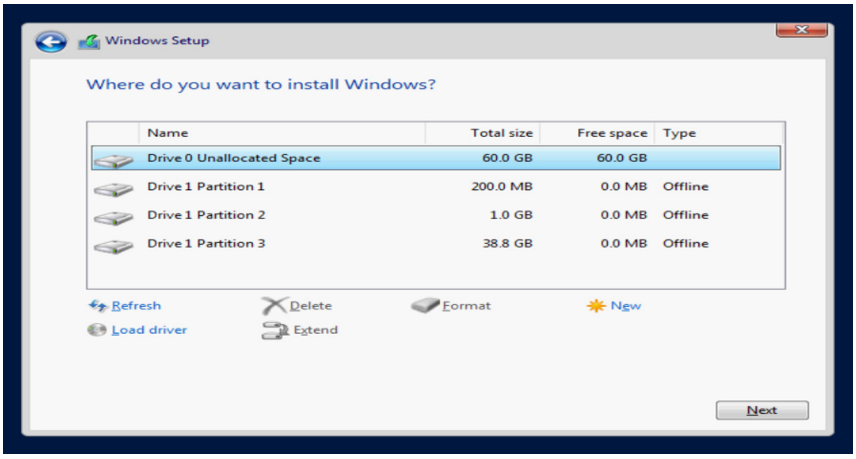


图 6-17. Windows 设置：选择安装目标

4. 在虚拟介质中安装驱动程序，以注入最新的 Marvell 驱动程序：

   a. 单击 **Load driver** （加载驱动程序），然后单击 **Browse** （浏览）（参见图 6-18）。

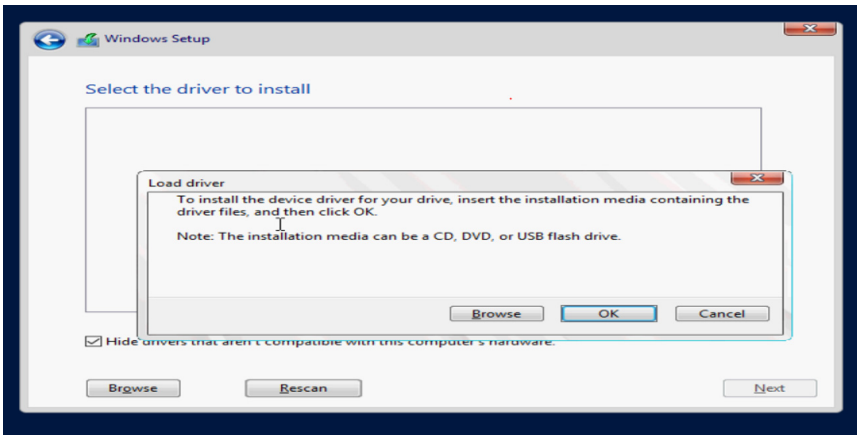


图 6-18. Windows 设置：选择要安装的驱动程序

   b. 导航到驱动程序的位置，并选择 qevbd 驱动程序。

   c. 选择要在其中安装驱动程序的适配器，然后单击 **Next** （下一步）继续。

5. 重复步骤 4 以加载 qeios 驱动程序 （Marvell L4 iSCSI 驱动程序）。

6. 在注入 qevbd 和 qeios 驱动程序之后，单击 **Next** （下一步）开始在 iSCSI LUN 上安装。然后按照屏幕说明进行操作。

   服务器在安装过程的一部分中将会多次重新引导，然后将从 SAN LUN 的 iSCSI 引导进行引导。

7. 若未自动引导，请访问 **Boot Menu** （引导菜单），然后选择特定的端口引导条目以从 iSCSI LUN 引导。

在 Linux 上配置从 SAN 的 iSCSI 引导

本节为以下 Linux 分发提供从 SAN 的 iSCSI 引导步骤：

- 从 RHEL 7.5 及更高版本的 SAN 配置 iSCSI 引导
- 从 SLES 12 SP3 及更高版本的 SAN 配置 iSCSI 引导
- 从 SAN 为其他 Linux 分发配置 iSCSI 引导

从 RHEL 7.5 及更高版本的 SAN 配置 iSCSI 引导

要安装 RHEL 7.5 及更高版本：

1. iSCSI 目标已在 UEFI 中连接时，从 RHEL 7.*x* 安装介质引导。

   ```
   Install Red Hat Enterprise Linux 7.x
   Test this media & install Red Hat Enterprise 7.x
   Troubleshooting -->

   Use the UP and DOWN keys to change the selection
   Press 'e' to edit the selected item or 'c' for a command
   prompt
   ```

2. 要安装开箱即用驱动程序，请按 E 键。否则，继续步骤 6。
3. 选择内核行，然后按 E 键。
4. 发出以下命令，然后按 ENTER。

   `inst.dd modprobe.blacklist=qed,qede,qedr,qedi,qedf`

   安装过程将提示您安装开箱即用的驱动程序。
5. 如果您的设置需要，请在提示其他驱动程序磁盘时加载 FastLinQ 驱动程序更新磁盘。或者，如果您没有其他驱动更新磁盘可安装，请按 C 键。
6. 继续安装。您可以跳过介质测试。单击 **Next** （下一步）继续安装。
7. 在 Configuration （配置）窗口中，选择在安装过程中要使用的语言，然后单击 **Continue** （继续）。
8. 在 Installation Summary （安装摘要）窗口中，单击 **Installation Destination** （安装目标）。磁盘标签是 *sda*，指示单路径安装。如果您配置了多路径，则该磁盘具有设备映射器标签。
9. 在 **Specialized & Network Disks** （专业化和网络磁盘）部分，选择 iSCSI LUN。
10. 键入根用户的密码，然后单击 **Next** （下一步）完成安装。
11. 重新引导服务器，然后在命令行中添加以下参数：

    `rd.iscsi.firmware`

    `rd.driver.pre=qed,qedi` （以加载所有驱动程序 `pre=qed,qedi,qede,qedf`）

    `selinux=0`
12. 成功引导系统后，编辑 `/etc/modprobe.d/anaconda-blacklist.conf` 文件，以删除所选驱动程序的黑名单条目。

13. 如下所示编辑 `/etc/default/grub` 文件：

    a. 找到双引号中的字符串，如以下示例所示。命令行是具体参考，有助于找到该字符串。

    ```
    GRUB_CMDLINE_LINUX="iscsi_firmware"
    ```

    b. 命令行可能包含保持不变的其他参数。只更改 iscsi_firmware 字符串，如下所示：

    ```
    GRUB_CMDLINE_LINUX="rd.iscsi.firmware selinux=0"
    ```

14. 发出以下命令，创建新的 `grub.cfg` 文件：

    ```
    # grub2-mkconfig -o /boot/efi/EFI/redhat/grub.cfg
    ```

15. 发出 **`dracut -f`** 命令重建虚拟内存盘，然后重新引导。

注

当在多路径 I/O (MPIO) 配置和单路径处于活动状态的 Linux 中安装 iSCSI·BFS 时，请在多路径 . conf 文件中使用以下设置：

```
defaults {
        find_multipaths yes
        user_friendly_names yes
        polling_interval 5
        fast_io_fail_tmo 5
        dev_loss_tmo 10
        checker_timeout 15
        no_path_retry queue
}
```

这些建议的设置可进行调整，并为 iSCSI·BFS 的操作提供指导。

有关更多信息，请联系相应的操作系统供应商。

从 SLES 12 SP3 及更高版本的 SAN 配置 iSCSI 引导

要安装 SLES 12 SP3 及更高版本：

1. 从 iSCSI 目标已预配置并且在 UEFI 中连接的 SLES 12 SP3 安装介质引导。
2. 在安装程序命令参数中添加 `dud=1` 参数，以更新最新的驱动程序包。需要驱动程序更新磁盘，因为必要的 iSCSI 驱动程序未内建。

注

仅适用于 SLES 12 SP3：如果服务器配置用于多功能模式 (NPAR)，则在此步骤中必须提供以下附加参数：

```
dud=1 brokenmodules=qed,qedi,qedf,qede withiscsi=1
[BOOT_IMAGE=/boot/x86_64/loader/linux dud=1
brokenmodules=qed,qedi,qedf,qede withiscsi=1]
```

3. 完成 SLES 12 SP3 OS 指定的安装步骤。

DHCP 配置中的已知问题

在 SLES 12 SP3 及更高版本的 DHCP 配置中，如果从 DHCP 服务器获取的启动器 IP 地址所处范围与目标 IP 地址不同， OS 安装后的第一引导可能会失败。要解决此问题，请使用静态配置引导 OS，更新最新的 iscsiuio 开箱即用 RPM，重建 initrd，然后使用 DHCP 配置重新引导 OS。 OS 现在应可成功引导。

从 SAN 为其他 Linux 分发配置 iSCSI 引导

对于 RHEL 6.9/6.10/7.2/7.3、 SLES 11 SP4 和 SLES 12 SP1/2 等分发版，内建 iSCSI 用户空间公用程序 （Open-iSCSI 工具）不支持 qedi iSCSI 传输，无法执行用户空间启动的 iSCSI 功能。从 SAN 引导安装期间，您可以使用驱动程序更新磁盘 (DUD) 来更新 qedi 驱动程序。不过，不存在用于更新用户空间内建公用程序的接口或进程，这会导致 iSCSI 目标登录和从 SAN 引导安装失败。

要克服这一限制，在从 SAN 引导期间，通过以下步骤之一使用纯 L2 接口 （不使用硬件卸载 iSCSI）执行从 SAN 的初始引导。

用于 Linux 其他分发版的 iSCSI 卸载包含以下信息：

- 使用软件启动器从 SAN 引导：
- 从软件 iSCSI 安装迁移到卸载 iSCSI
- Linux 多路径注意事项

使用软件启动器从 SAN 引导：

用 Dell OEM 解决方案使用软件启动器从 SAN 引导：

注

上述步骤是必需的，因为 DUD 包含 qedi，而后者绑定至 iSCSI PF。绑定后，由于传输驱动程序未知，因此 Open-iSCSI 基础架构失败。

1. 访问 Dell EMC 系统 BIOS 设置。
2. 配置启动器和目标条目。有关更多信息，请参阅 Dell BIOS 用户指南。
3. 在安装开始时，通过 DUD 选项传递以下引导参数：
   - ❑ 对于 RHEL 6.*x*、 7.*x* 及更早版本：

     `rd.iscsi.ibft dd`

     较旧分发版的 RHEL 无需单独的选项。
   - ❑ 对于 SLES 11 SP4 和 SLES 12 SP1/SP2：

     `ip=ibft dud=1`
   - ❑ 对于 FastLinQ DUD 包 （例如，在 RHEL 7 中）：

     `fastlinq-8.18.10.0-dd-rhel7u3-3.10.0_514.el7-x86_64.iso`

     其中 RHEL 7.*x* 的 DUD 参数是 `dd`， SLES 12.*x* 的是 `dud=1`。
4. 在目标 LUN 上安装 OS。

从软件 iSCSI 安装迁移到卸载 iSCSI

按照您操作系统的说明，从非卸载界面迁移到卸载界面。

- 对 RHEL 6.9/6.10 迁移到卸载 iSCSI
- 对 SLES 11 SP4 迁移到卸载 iSCSI
- 对 SLES 12 SP1/SP2 迁移到卸载 iSCSI

对 RHEL 6.9/6.10 迁移到卸载 iSCSI

要对 RHEL 6.9 或 6.10 从软件 iSCSI 安装迁移到卸载 iSCSI：

1. 引导到 iSCSI 非卸载 / 从 SAN 操作系统引导的 L2。发出以下命令来安装 Open-iSCSI 和 iscsiuio RPM：

```
# rpm -ivh --force qlgc-open-iscsi-2.0_873.111-1.x86_64.rpm
# rpm -ivh --force iscsiuio-2.11.5.2-1.rhel6u9.x86_64.rpm
```

注

要在安装时保留内建 RPM 的原始内容，必须使用 `-ivh` 选项（而非 `-Uvh` 选项），然后使用 `--force` 选项。

2. 编辑 `/etc/init.d/iscsid` 文件，添加以下命令，然后保存文件：

```
modprobe -q qedi
```

例如：

```
echo -n $"Starting $prog: "
modprobe -q iscsi_tcp
modprobe -q ib_iser
modprobe -q cxgb3i
modprobe -q cxgb4i
modprobe -q bnx2i
modprobe -q be2iscsi
modprobe -q qedi
daemon iscsiuio
```

3. 打开 `/boot/efi/EFI/redhat/grub.conf` 文件，进行以下更改并保存该文件：

- 删除 `ifname=eth5:14:02:ec:ce:dc:6d`
- 删除 `ip=ibft`
- 添加 `selinux=0`

例如：

```
kernel /vmlinuz-2.6.32-696.el6.x86_64 ro
root=/dev/mapper/vg_prebooteit-lv_root rd_NO_LUKS
iscsi_firmware LANG=en_US.UTF-8 ifname=eth5:14:02:ec:ce:dc:6d
rd_NO_MD SYSFONT=latarcyrheb-sun16 crashkernel=auto rd_NO_DM
rd_LVM_LV=vg_prebooteit/lv_swap ip=ibft  KEYBOARDTYPE=pc
KEYTABLE=us rd_LVM_LV=vg_prebooteit/lv_root rhgb quiet
        initrd /initramfs-2.6.32-696.el6.x86_64.img

kernel /vmlinuz-2.6.32-696.el6.x86_64 ro
root=/dev/mapper/vg_prebooteit-lv_root rd_NO_LUKS
iscsi_firmware LANG=en_US.UTF-8 rd_NO_MD
SYSFONT=latarcyrheb-sun16 crashkernel=auto rd_NO_DM
rd_LVM_LV=vg_prebooteit/lv_swap KEYBOARDTYPE=pc KEYTABLE=us
rd_LVM_LV=vg_prebooteit/lv_root selinux=0
        initrd /initramfs-2.6.32-696.el6.x86_64.img
```

4. 发出以下命令，以建立 `initramfs` 文件：

```
# dracut -f
```

5. 重新引导服务器，然后打开 UEFI HII。
6. 在 HII 中，禁用 iSCSI 从 BIOS 引导，然后对适配器启用 iSCSI HBA 或引导，如下所示：
   a. 选择适配器端口，然后选择 **Device Level Configuration** （设备级配置）。
   b. 在 Device Level Configuration （设备级配置）页面上，为 **Virtualization Mode** （虚拟化模式）选择 **NPAR**。
   c. 打开 NIC Partitioning Configuration （NIC 分区配置）页面，将 **iSCSI Offload Mode** （iSCSI 卸载模式）设置为 **Enabled** （已启用）。（iSCSI HBA 支持在双端口适配器的分区 3 上和四端口适配器的分区 2 上。）
   d. 打开 **NIC Configuration** （NIC 配置）菜单，将 **Boot Protocol** （引导协议）设置为 **UEFI iSCSI**。
   e. 打开 iSCSI Configuration （iSCSI 配置）页面并配置 iSCSI 设置。
7. 保存配置并重新引导服务器。

现在，OS 可通过卸载接口引导。

对 SLES 11 SP4 迁移到卸载 iSCSI

要对 SLES 11 SP4 从软件 iSCSI 安装迁移到卸载 iSCSI：

1. 发出以下命令将 Open-iSCSI 工具和 iscsiuio 更新到最新可用版本：

```
# rpm -ivh qlgc-open-iscsi-2.0_873.111.sles11sp4-3.x86_64.rpm --force
# rpm -ivh iscsiuio-2.11.5.3-2.sles11sp4.x86_64.rpm --force
```

注

要在安装时保留内建 RPM 的原始内容，必须使用 `-ivh` 选项 （而非 `-Uvh` 选项），然后使用 `--force` 选项。

2. 编辑 `/etc/elilo.conf` 文件，进行以下更改，然后保存该文件：
   - ❑ 删除 `ip=ibft` 参数 (如果存在)
   - ❑ 添加 `iscsi_firmware`

3. 如下所示编辑 /etc/sysconfig/kernel 文件：

   a. 找到以 INITRD_MODULES 开头的行。此行看起来类似如下，但可能包含不同的参数：

      ```
      INITRD_MODULES="ata_piix ata_generic"
      ```

      或者

      ```
      INITRD_MODULES="ahci"
      ```

   b. 在现有行末（引号内）添加 qedi 以编辑该行。例如：

      ```
      INITRD_MODULES="ata_piix ata_generic qedi"
      ```

      或者

      ```
      INITRD_MODULES="ahci qedi"
      ```

   c. 保存文件。

4. 编辑 /etc/modprobe.d/unsupported-modules 文件，将 allow_unsupported_modules 的值更改为 1，然后保存该文件：

   ```
   allow_unsupported_modules 1
   ```

5. 找到并删除以下文件：

   - /etc/init.d/boot.d/K01boot.open-iscsi
   - /etc/init.d/boot.open-iscsi

6. 创建 initrd 的备份，然后通过发出以下命令来构建新的 initrd。

   ```
   # cd /boot/
   # mkinitrd
   ```

7. 重新引导服务器，然后打开 UEFI HII。

8. 在 UEFI HII 中，禁用 iSCSI 从 BIOS 引导，然后对适配器启用 iSCSI HBA 或引导，如下所示：

   a. 选择适配器端口，然后选择 **Device Level Configuration**（设备级配置）。

   b. 在 Device Level Configuration（设备级配置）页面上，为 **Virtualization Mode**（虚拟化模式）选择 **NPAR**。

   c. 打开 NIC Partitioning Configuration（NIC 分区配置）页面，将 **iSCSI Offload Mode**（iSCSI 卸载模式）设置为 **Enabled**（已启用）。（iSCSI HBA 支持在双端口适配器的分区 3 上和四端口适配器的分区 2 上。）

   d. 打开 **NIC Configuration**（NIC 配置）菜单，将 **Boot Protocol**（引导协议）设置为 **UEFI iSCSI**。

e. 打开 iSCSI Configuration （iSCSI 配置）页面并配置 iSCSI 设置。

9. 保存配置并重新引导服务器。

现在，OS 可通过卸载接口引导。

对 SLES 12 SP1/SP2 迁移到卸载 iSCSI

要对 SLES 12 SP1/SP2 从软件 iSCSI 安装迁移到卸载 iSCSI：

1. 引导到 iSCSI 非卸载 / 从 SAN 操作系统引导的 L2。发出以下命令来安装 Open-iSCSI 和 iscsiuio RPM：

```
# rpm -ivh qlgc-open-iscsi-2.0_873.111.sles12p2-3.x86_64.rpm --force
# rpm -ivh iscsiuio-2.11.5.3-2.sles12sp2.x86_64.rpm --force
```

> **注**
>
> 要在安装时保留内建 RPM 的原始内容，必须使用 `-ivh` 选项 （而非 `-Uvh` 选项），然后使用 `--force` 选项。

2. 通过发出以下命令重新加载所有守护进程服务：

```
# systemctl daemon-reload
```

3. 如果尚未通过发出以下命令启用 iscsid 和 iscsiuio 服务，请启用 iscsid 和 iscsiuio 服务：

```
# systemctl enable iscsid
# systemctl enable iscsiuio
```

4. 发出以下命令：

```
cat /proc/cmdline
```

5. 检查 OS 是否保留了任何引导选项，例如 `ip=ibft` 或 `rd.iscsi.ibft`。
   - ❑ 如果有保留的引导选项，继续步骤 6。
   - ❑ 如果没有保留的引导选项，跳至步骤 6c。
6. 编辑 `/etc/default/grub` 文件并修改 `GRUB_CMDLINE_LINUX` 值：
   a. 移除 `rd.iscsi.ibft` （如果存在）。
   b. 移除任何 `ip=<value>` 引导选项 （如果存在）。
   c. 添加 `rd.iscsi.firmware`。对于较老的发行版，添加 `iscsi_firmware`。

7. 创建原文件 grub.cfg 的备份。该文件位于以下位置：
   - 传统引导：/boot/grub2/grub.cfg
   - UEFI 引导：SLES 的 /boot/efi/EFI/sles/grub.cfg
8. 发出以下命令，创建新的 grub.cfg 文件：

   ```
   # grub2-mkconfig -o <new file name>
   ```
9. 将旧 grub.cfg 文件与新 grub.cfg 文件进行比较，以验证更改。
10. 将原始 grub.cfg 文件替换为新 grub.cfg 文件。
11. 通过发出以下命令建立 initramfs 文件：

    ```
    # dracut -f
    ```
12. 重新引导服务器，然后打开 UEFI HII。
13. 在 UEFI HII 中，禁用 iSCSI 从 BIOS 引导，然后对适配器启用 iSCSI HBA 或引导，如下所示：
    a. 选择适配器端口，然后选择 **Device Level Configuration** （设备级配置）。
    b. 在 Device Level Configuration （设备级配置）页面上，为 **Virtualization Mode** （虚拟化模式）选择 **NPAR**。
    c. 打开 NIC Partitioning Configuration （NIC 分区配置）页面，将 **iSCSI Offload Mode** （iSCSI 卸载模式）设置为 **Enabled** （已启用）。（iSCSI HBA 支持在双端口适配器的分区 3 上和四端口适配器的分区 2 上。）
    d. 打开 **NIC Configuration** （NIC 配置）菜单，将 **Boot Protocol** （引导协议）设置为 **UEFI iSCSI**。
    e. 打开 iSCSI Configuration （iSCSI 配置）页面并配置 iSCSI 设置。
14. 保存配置并重新引导服务器。

现在，OS 可通过卸载接口引导。

Linux 多路径注意事项

Linux 操作系统上的 iSCSI 从 SAN 引导安装目前仅支持单路径配置。要启用多路径配置，请使用 L2 或 L4 路径以单一路径执行安装。在服务器引导到安装的操作系统后，执行必要的配置以启用多路径 I/O (MPIO)。

请参阅本节中适当的步骤，以从 L2 迁移到 L4，并为您的 OS 配置 MPIO：

- 对 RHEL 6.9/6.10 迁移并配置 MPIO 到卸载的界面
- 对 SLES 11 SP4 迁移并配置 MPIO 到卸载的界面
- 对 SLES 12 SP1/SP2 迁移并配置 MPIO 到卸载的界面

对 RHEL 6.9/6.10 迁移并配置 MPIO 到卸载的界面

要对 RHEL 6.9/6.10 从 L2 迁移到 L4 并配置 MPIO 引导 OS 到卸载的界面：

1. 在适配器的双端口上配置 L2 BFS 的 iSCSI 引导设置。引导只通过一个端口登录，但会为两个端口创建 iSCSI 引导固件表 (iBFT) 界面。
2. 在引导到 CD 时，确保指定以下内核参数：

   ```
   ip=ibft
   linux dd
   ```

3. 提供 DUD 并完成安装。
4. 通过 L2 引导到 OS。
5. 通过发出以下命令更新 Open-iSCSI 工具和 iscsiuio：

```
# rpm -ivh qlgc-open-iscsi-2.0_873.111.rhel6u9-3.x86_64.rpm --force
# rpm -ivh iscsiuio-2.11.5.5-6.rhel6u9.x86_64.rpm --force
```

6. 编辑 `/boot/efi/EFI/redhat/grub.conf` 文件，进行以下更改，然后保存该文件：
   a. 删除 `ifname=eth5:14:02:ec:ce:dc:6d`。
   b. 删除 `ip=ibft`。
   c. 添加 `selinux=0`。
7. 发出以下命令以建立 `initramfs` 文件：

   ```
   # dracut -f
   ```

8. 重新引导，并更改适配器引导设置以对双端口使用 L4 或 **iSCSI (HW)**，以及通过 L4 引导。

9. 在引导到 OS 后，设置多路径守护进程 `multipathd.conf`：

```
# mpathconf --enable --with_multipathd y
# mpathconf -enable
```

10. 启动多路径服务：

```
# service multipathd start
```

11. 重建支持多路径的 `initramfs`。

```
# dracut --force --add multipath --include /etc/multipath
```

12. 重新引导服务器并通过多路径引导到 OS。

注

为使 `/etc/multipath.conf` 文件中的任何其他更改生效，必须重建 initrd 映像并重新引导服务器。

对 SLES 11 SP4 迁移并配置 MPIO 到卸载的界面

要对 SLES 11 SP4 从 L2 迁移到 L4 并配置 MPIO 引导 OS 到卸载的界面：

1. 执行所有必要的步骤，通过单一路径将非卸载 (L2) 界面迁移到卸载 (L4) 界面。请参阅 对 SLES 11 SP4 迁移到卸载 iSCSI。
2. 在使用 L4 界面引导到 OS 后，如下所示准备多路径：
   a. 重新引导服务器，转到 **System Setup** （系统设置） **/Utilities** （公用程序）打开 HII。
   b. 在 HII 中，选择 **System Configuration** （系统配置），然后选择要用于多路径的第二个适配器端口。
   c. 在主要配置页面 （Main Configuration Page）的 **Port Level Configuration** （端口级配置）下，将 **Boot Mode** （引导模式）设置为 **iSCSI (HW)** 并启用 **iSCSI Offload** （iSCSI 卸载）。
   d. 在主要配置页面 （Main Configuration Page）的 **iSCSI Configuration** （iSCSI 配置）下，执行必要的 iSCSI 配置。
   e. 重新引导服务器并径引导到 OS。
3. 设置 MPIO 服务在重启时保持不变，如下所示：

```
# chkconfig multipathd on
# chkconfig boot.multipath on
# chkconfig boot.udev on
```

4. 启动多路径服务，如下所示：

```
# /etc/init.d/boot.multipath start
# /etc/init.d/multipathd start
```

5. 运行 `multipath -v2 -d` 以在干运行时显示多路径配置。
6. 查找 `/etc/multipath.conf` 下的 `multipath.conf` 文件。

注

如果该文件不存在，则从 `/usr/share/doc/packages/multipath-tools` 文件夹复制 `multipath.conf`：

```
cp /usr/share/doc/packages/multipath-tools/multipath.
conf.defaults /etc/multipath.conf
```

7. 编辑 `multipath.conf` 以启用默认部分。
8. 重建 initrd 映像以包含 MPIO 支持：

```
# mkinitrd -f multipath
```

9. 重新引导服务器并通过多路径支持引导 OS。

注

为使 `/etc/multipath.conf` 文件中的任何其他更改生效，必须重建 initrd 映像并重新引导服务器。

对 SLES 12 SP1/SP2 迁移并配置 MPIO 到卸载的界面

要对 SLES 12 SP1/SP2 从 L2 迁移到 L4 并配置 MPIO 引导 OS 到卸载的界面：

1. 在适配器的双端口上配置 L2 BFS 的 iSCSI 引导设置。引导只通过一个端口登录，但会为两个端口创建 iSCSI 引导固件表 (iBFT) 界面。
2. 在引导到 CD 时，确保指定以下内核参数：

```
dud=1
rd.iscsi.ibft
```

3. 提供 DUD 并完成安装。
4. 通过 L2 引导到 OS。

5. 通过发出以下命令更新 Open-iSCSI 工具：

```
# rpm -ivh qlgc-open-iscsi-2.0_873.111.sles12sp1-3.x86_64.rpm --force
# rpm -ivh iscsiuio-2.11.5.5-6.sles12sp1.x86_64.rpm --force
```

6. 将 `rd.iscsi.ibft` 参数更改为 `rd.iscsi.firmware`，编辑 `/etc/default/grub` 文件，然后保存该文件。

7. 发出以下命令：

```
# grub2-mkconfig -o /boot/efi/EFI/suse/grub.cfg
```

8. 要加载多路径模块，请发出以下命令：

```
# modprobe dm_multipath
```

9. 要启用多路径守护进程，请发出以下命令：

```
# systemctl start multipathd.service
# systemctl enable multipathd.service
# systemctl start multipathd.socket
```

10. 要运行多路径公用程序，请发出以下命令：

```
# multipath （可能不会显示多路径设备，因为它使用 L2 上的单个路径引导）
# multipath -ll
```

11. 要在 initrd 中插入多路径模块，请发出以下命令：

```
# dracut --force --add multipath --include /etc/multipath
```

12. 重新引导服务器并通过在 POST 菜单中按 F9 键进入系统设置。

13. 更改 UEFI 配置以使用 L4 iSCSI 引导：

    a. 打开系统配置，选择适配器端口，然后选择 **Port Level Configuration**（端口级配置）。

    b. 在 Port Level Configuration （端口级配置）页面上，将 **Boot Mode**（引导模式）设置为 **iSCSI (HW)**，并将 **iSCSI Offload** （iSCSI 卸载）设置为 **Enabled** （已启用）。

    c. 保存设置，然后退出系统配置菜单。

14. 重新引导系统。现在，OS 应该通过卸载接口引导。

注

为使 `/etc/multipath.conf` 文件中的任何其他更改生效，必须重建 initrd 映像并重新引导服务器。

在 VMware 上配置 iSCSI 从 SAN 引导

由于 VMware 本身不支持 iSCSI 从 SAN 卸载引导，因此您必须通过软件在预引导中配置 BFS，然后在 OS 驱动程序加载时过渡到卸载。有关更多信息，请参阅第 69 页上 “ 启用 NPAR 和 iSCSI HBA”。

在 VMware ESXi 中， iSCSI BFS 配置步骤包括：

- 设置 UEFI 主配置
- 为 iSCSI 引导 (L2) 配置系统 BIOS
- 映射 OS 安装的 CD 或 DVD

设置 UEFI 主配置

在 VMware 上配置 iSCSI 从 SAN 引导：

1. 将 41*xxx* 系列适配器插入 Dell 14G 服务器。例如，将 PCIE 和 LOM （四端口或双端口）插入 R740 服务器。
2. 在 HII 中，转到 **System Setup** （系统设置），选择 **Device Settings** （设备设置），然后选择要配置的集成式 NIC 端口。单击 **Finish** （完成）。
3. 在 **Main Configuration Page** （主要配置页面）上，选择 **NIC Partitioning Configuration** （NIC 分区配置），然后单击 **Finish** （完成）。
4. 在 **Main Configuration Page** （主要配置页面）上，选择 **Firmware Image Properties** （固件映像属性），查看不可配置的信息，然后单击 **Finish** （完成）。
5. 在 **Main Configuration Page** （主要配置页面）上，选择 **Device Level Configuration** （设备级配置）。
6. 完成 Device Level Configuration （设备级配置）页面 （参见图 6-19），如下所示：
   a. 对于 **Virtualization Mode** （虚拟化模式），为通过 NIC 界面的 IBFT 安装选择 **None** （无）、 **NPar** 或 **NPar_EP**。
   b. 对于 **NParEP Mode** （NParEP 模式），选择 **Disabled** （已禁用）。

c. 为 **UEFI Driver Debug Level** （UEFI 驱动程序调试级别）选择 **10**。

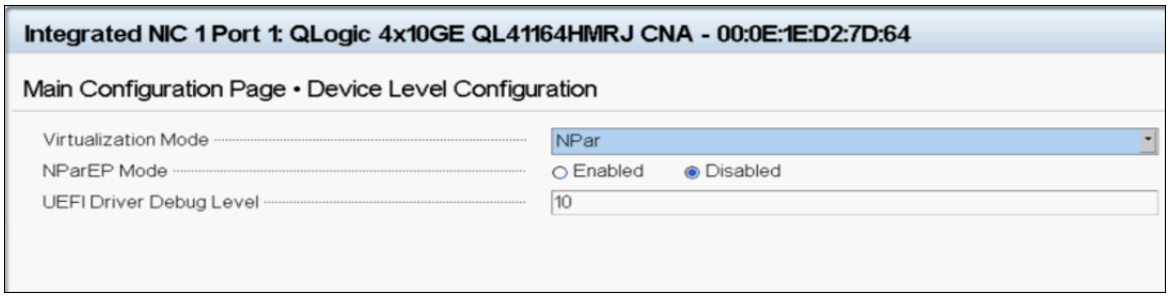


图 6-19. 集成式 NIC：VMware 的设备级配置

7. 进入 **Main Configuration Page** （主要配置页面），然后选择 **NIC Partitioning Configuration** （NIC 分区配置）。
8. 在 NIC Partitioning Configuration （NIC 分区配置）页面上，选择 **Partition 1 Configuration** （分区 1 配置）。
9. 完成 Partition 1 Configuration （分区 1 配置）页面，如下所示：
   a. 为 **Link Speed** （链路速度）选择 **Auto Neg** （自动协商）、 **10Gbps** 或 **1Gbps**。
   b. 确保链路可以运行。
   c. 为 **Boot Protocol** （引导协议）选择 **None** （无）。
   d. 对于 **Virtual LAN Mode** （虚拟 LAN 模式），选择 **Disabled** （已禁用）。
10. 在 NIC Partitioning Configuration （NIC 分区配置）页面上，选择 **Partition 2 Configuration** （分区 2 配置）。
11. 完成 Partition 2 Configuration （分区 2 配置）页面（参见图 6-20），如下所示：
    a. 对于 **FCoE Mode** （FCoE 模式），选择 **Disabled** （已禁用）。
    b. 对于 **iSCSI Offload Mode** （iSCSI 卸载模式），选择 **Disabled** （已禁用）。

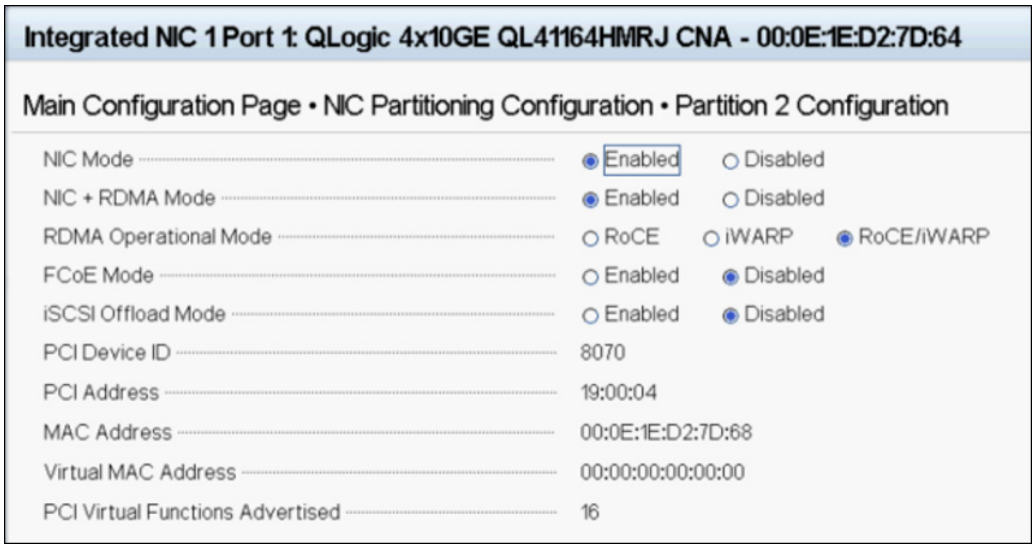


图 6-20. 集成式 NIC：VMware 的分区 2 配置

为 iSCSI 引导 (L2) 配置系统 BIOS

要在 VMware 上配置系统 BIOS：

1. 在 System BIOS Settings （系统 BIOS 设置）页面上，单击 **Boot Settings**（引导设置）。
2. 如图 6-21 所示完成 Boot Settings （引导设置）页面。

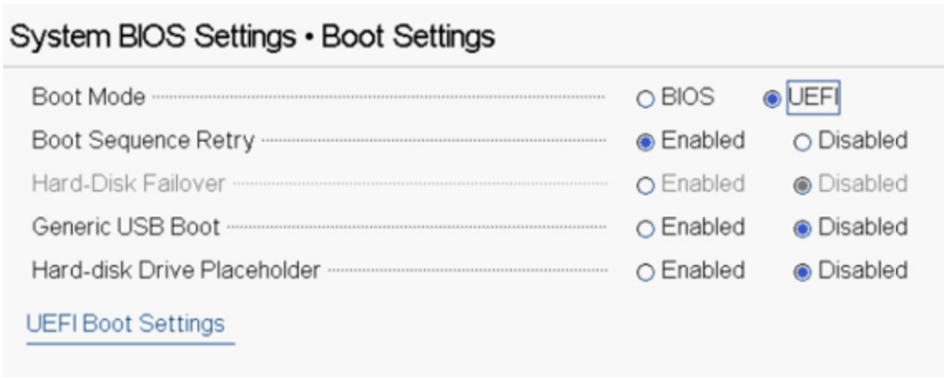


图 6-21. 集成式 NIC：系统 BIOS， VMware 的引导设置

3. 在 System BIOS Settings （系统 BIOS 设置）页面上，选择 **Network Settings** （网络设置）。

4. 在 Network Settings （网络设置）页面的 **UEFI iSCSI Settings** （UEFI iSCSI 设置）下：
   a. 为 **iSCSI Device1** （iSCSI 设备 1）选择 **Enabled** （已启用）。
   b. 选择 **UEFI Boot Settings** （UEFI 引导设置）。
5. 在 iSCSI Device1 （iSCSI 设备 1）页面上：
   a. 为 **Connection 1** （连接 1）选择 **Enabled** （已启用）。
   b. 选择 **Connection 1 Settings** （连接 1 设置）。
6. 在 Connection 1 Settings （连接 1 设置）页面 （参见图 6-22）上：
   a. 对于 **Interface** （接口），选择要在其上测试 iSCSI 引导固件表 (IBFT) 从 SAN 引导的适配器端口。
   b. 对于 **Protocol** （协议），选择 **Ipv4** 或 **IPv6**。
   c. 对于 **VLAN**，选择 **Disabled** （已禁用）（默认值）或 **Enabled** （已启用）（如果要使用 vLAN 测试 IBFT）。
   d. 对于 **DHCP**，选择 **Enabled** （已启用）（如果 IP 地址来自 DHCP 服务器）或 **Disabled** （已禁用）（以使用静态 IP 配置）。
   e. 对于 **Target info via DHCP** （通过 DHCP 的目标信息），选择 **Disabled** （已启用）。

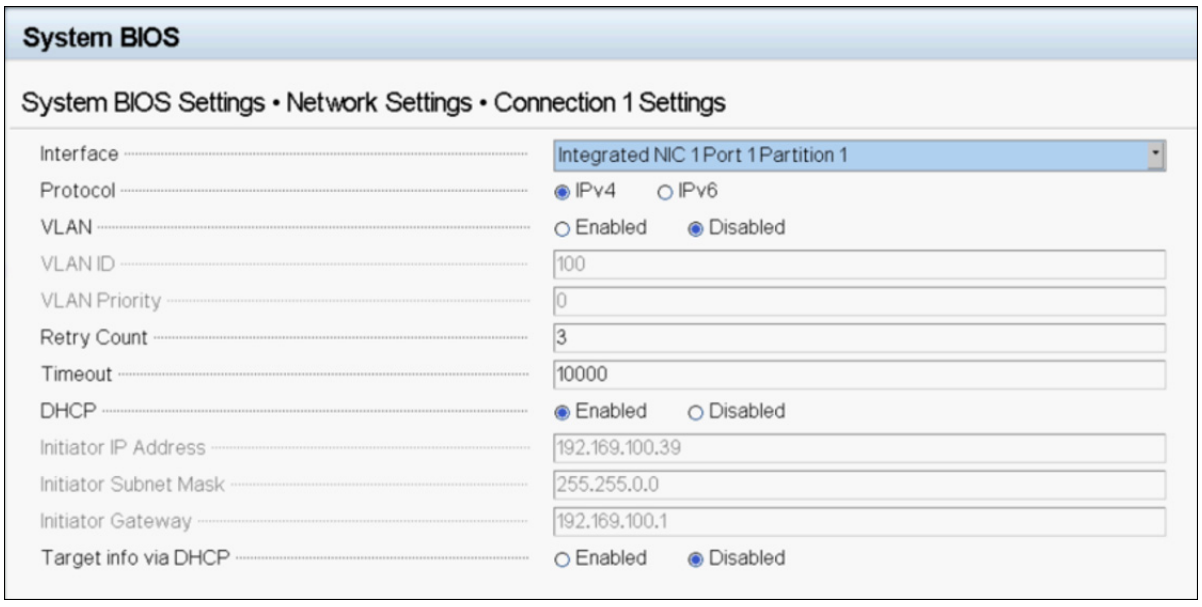


图 6-22. 集成式 NIC：系统 BIOS， VMware 的连接 1 设置

7. 完成目标详细信息，并且为 **Authentication Type** （身份验证类型）选择 **CHAP** （以设置 CHAP 详细信息）或 **None** （无）（默认值）。图 6-23 显示一个示例。

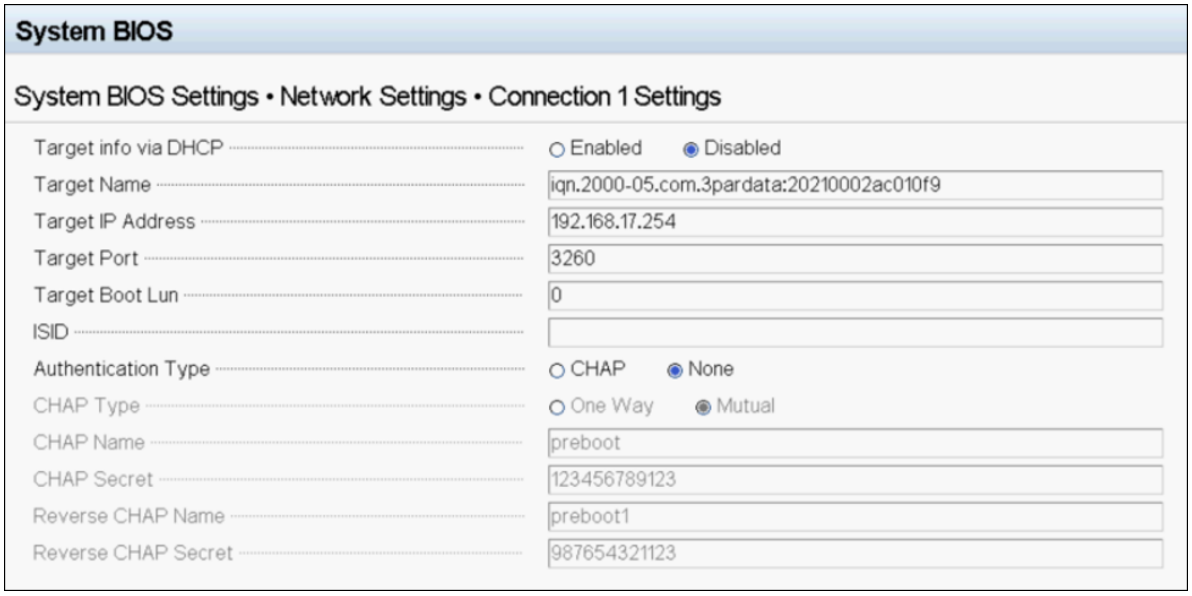


图 6-23. 集成式 NIC：系统 BIOS， VMware 的连接 1 设置 （目标）

8. 保存所有配置更改，然后重新引导服务器。
9. 在系统启动时，按 F11 键引导引导管理器。
10. 在引导管理器的 **Boot Menu** （引导菜单）、 **Select UEFI Boot Option** （选择 UEFI 引导选项）下，选择 **Embedded SATA Port AHCI Controller** （嵌入式 SATA 端口 AHCI 控制器）。

映射 OS 安装的 CD 或 DVD

要映射 CD 或 DVD：

1. 使用 ESXi-Customizer 创建自定义的 ISO 映像，并插入最新的驱动程序包或 VIB。
2. 将 ISO 映射到服务器虚拟控制台的虚拟介质。
3. 在虚拟光盘上，加载 ISO 文件。
4. 在 ISO 成功加载后，按 F11 键。

5. 在 Select a Disk To Install Or Upgrade （选择要安装或升级的磁盘）窗口的 **Storage Device**（存储设备）下，选择 **3PARdata W** 磁盘，然后按 ENTER 键。图 6-24 显示一个示例。

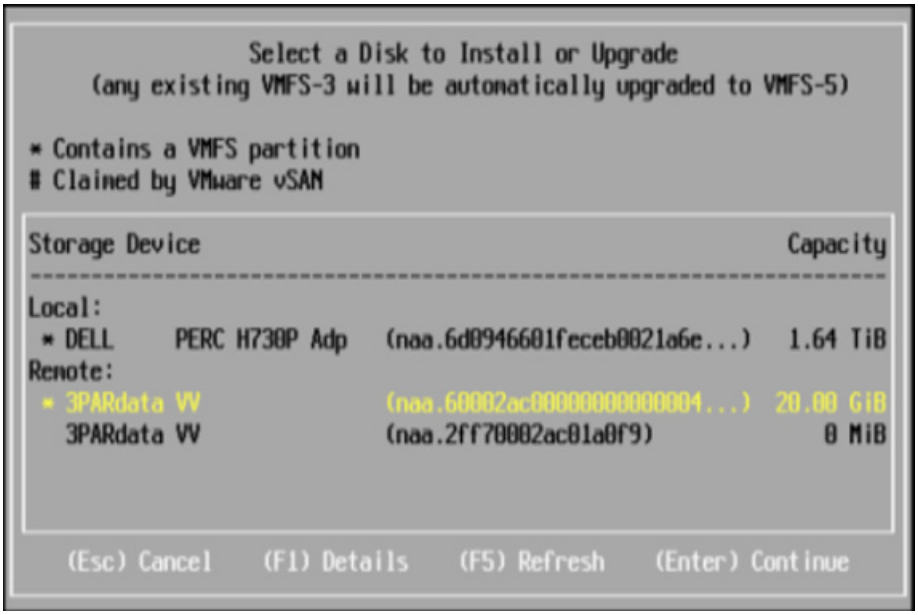


图 6-24. VMware iSCSI BFS：选择要安装的磁盘

6. 在远程 iSCSI LUN 上启动 ESXi OS 的安装。
7. 在 ESXi OS 安装成功完成后，系统将引导到 OS，如图 6-25 所示。

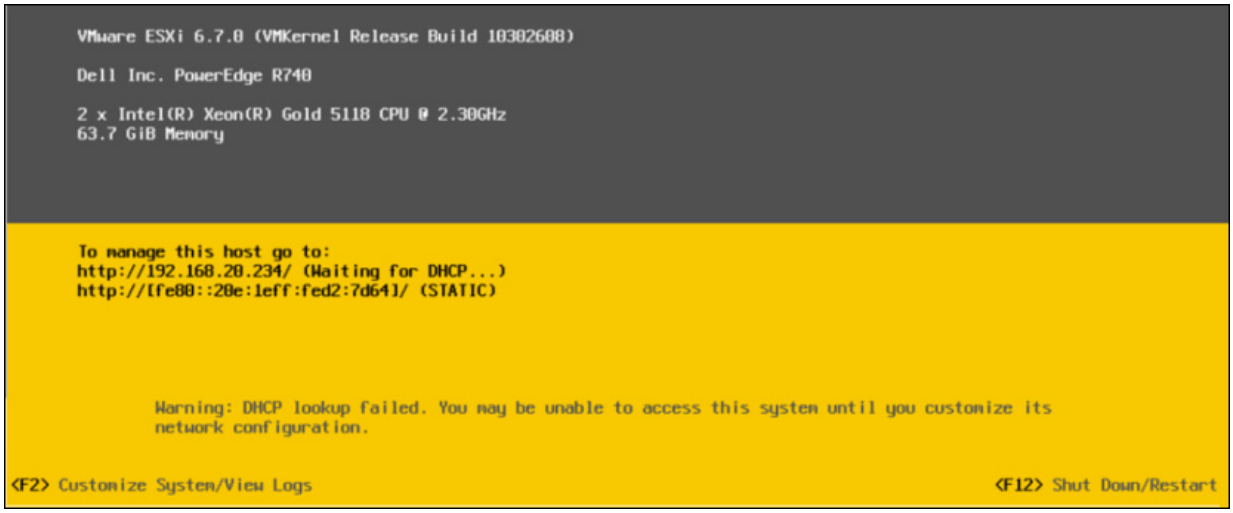


图 6-25. VMware iSCSI 从 SAN 引导成功

从 SAN 的 FCoE 引导

Marvell 41*xxx* 系列适配器 支持 FCoE 引导以使操作系统网络引导到无磁盘系统。FCoE 引导允许 Windows、Linux 或 VMware 操作系统通过 FCoE 支持网络从位于远程的光纤通道或 FCoE 目标计算机引导。您可以打开 **NIC Configuration**（NIC 配置）菜单并将 **Boot Protocol**（引导协议）选项设置为 **FCoE**，以设置 FCoE 选项 （使用 Marvell 卸载 FCoE 驱动程序的卸载路径）。

本节提供有关从 SAN 的 FCoE 引导的以下配置信息：

- FCoE 开箱即用和内建支持
- FCoE 预引导配置
- 在 Windows 上配置从 SAN 的 FCoE 引导
- 在 Linux 上配置从 SAN 的 FCoE 引导
- 在 VMware 上配置 SAN 的 FCoE 引导

FCoE 开箱即用和内建支持

表 6-6 列出操作系统的内建和开箱即用从 SAN 的 FCoE 引导 (BFS) 支持。

表 6-6. FCoE 开箱即用和内建从 SAN 引导支持

OS 版本	开箱即用 硬件卸载 FCoE BFS 支持	内建 硬件卸载 FCoE BFS 支持
Windows 2012	是	否
Windows 2012 R2	是	否
Windows 2016	是	否
Windows 2019	是	是
RHEL 7.5	是	是
RHEL 7.6	是	是
RHEL 8.0	是	是
SLES 15/15 SP1	是	是
vSphere ESXi 6.5 U3	是	否
vSphere ESXi 6.7 U2	是	否

FCoE 预引导配置

本节介绍 Windows、 Linux 和 ESXi 操作系统的安装和引导步骤。要准备系统 BIOS，请修改系统引导顺序并指定 BIOS 引导协议 （如果需要）。

注

ESXi 5.5 及更高版本支持从 SAN 的 FCoE 引导。并非所有适配器版本均支持 FCoE 和从 SAN 的 FCoE 引导。

指定 BIOS 引导协议

从 SAN 的 FCoE 引导仅在 UEFI 模式下受支持。使用系统 BIOS 配置将引导模式（协议）中的平台设置为 UEFI。

注

FCoE BFS 在旧版 BIOS 模式下不受支持。

配置适配器 UEFI 引导模式

要将引导模式配置为 FCOE：

1. 重新启动系统。
2. 按 OEM 热键进入 System Setup （系统设置）（图 6-26）。这也称为 UEFI HII。

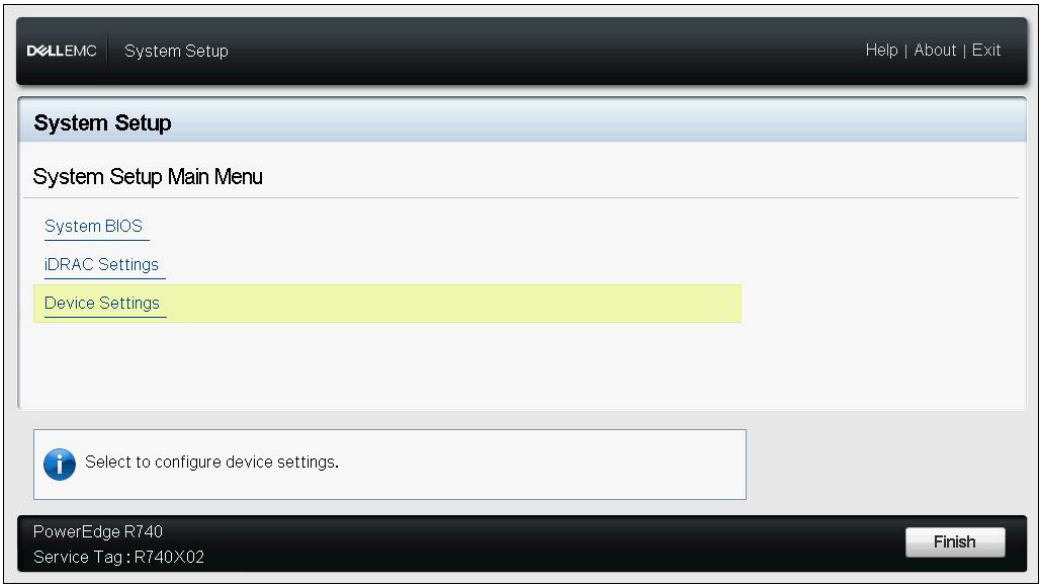


图 6-26. 系统设置：选择设备设置

注

SAN 引导仅在 UEFI 环境中受支持。确保系统引导选项为 UEFI，而非旧版。

3. 在 Device Settings （设备设置）页面上，选择 Marvell FastLinQ 适配器（图 6-27）。

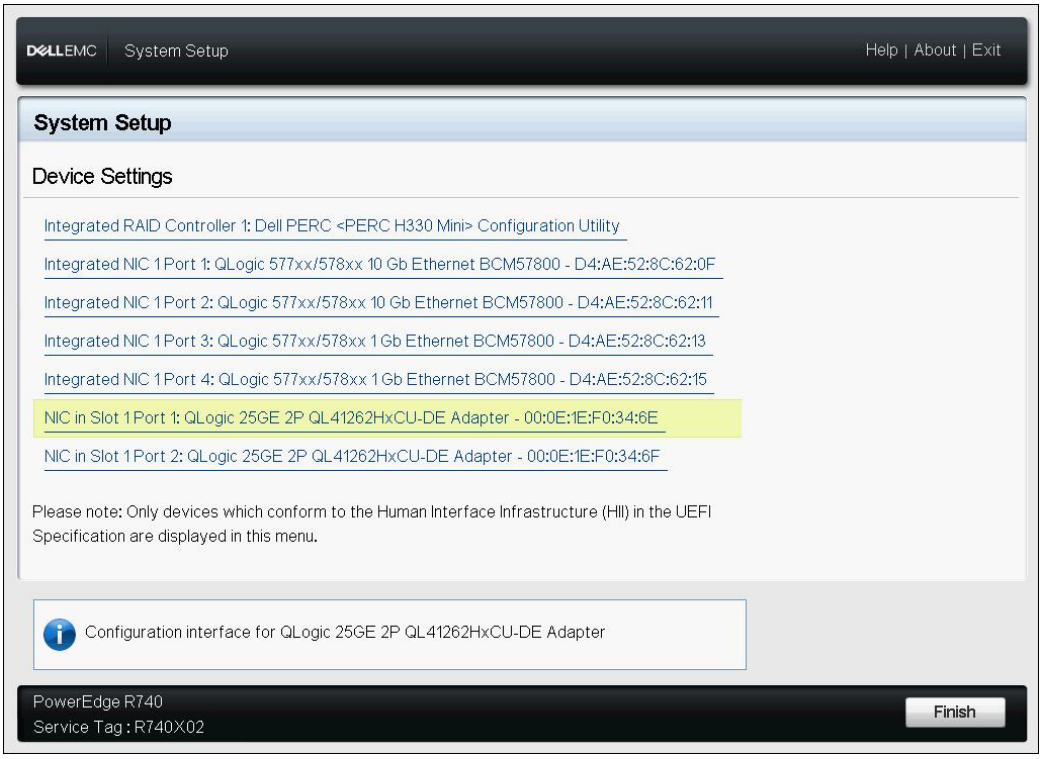


图 6-27. 系统设置：设备设置、端口选择

4. 在 Main Configuration Page （主要配置页面）上，选择 **NIC Configuration**（NIC 配置）（图 6-28），然后按 ENTER 键。

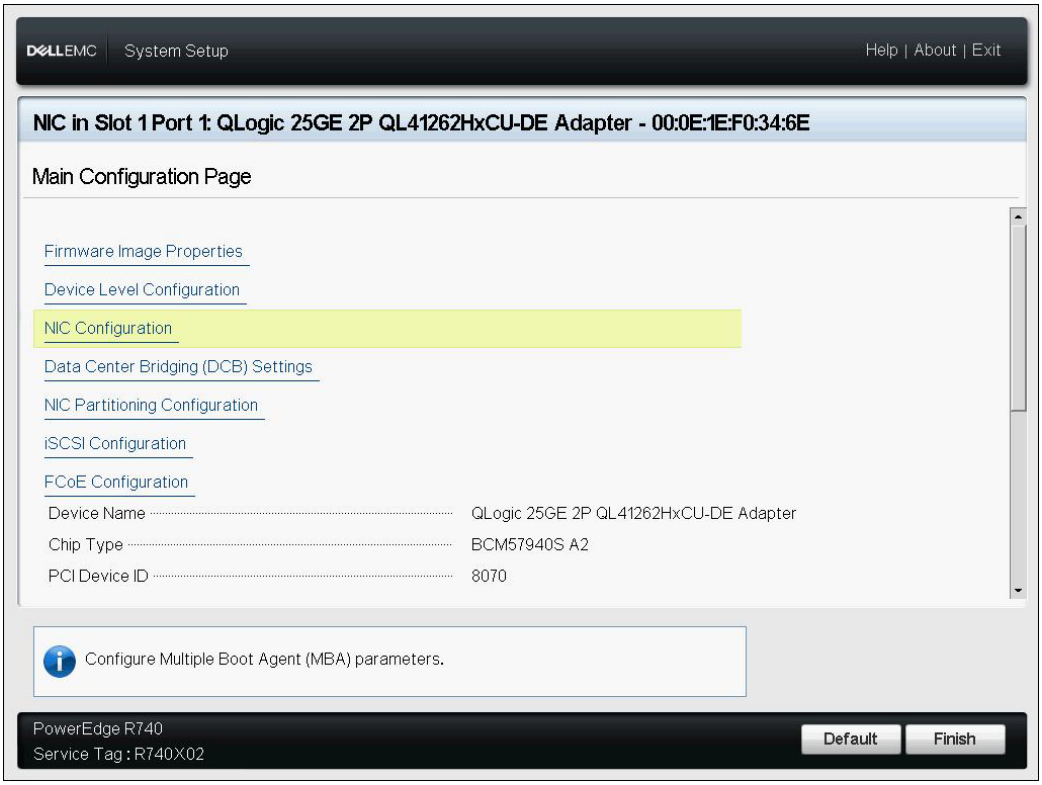


图 6-28. 系统设置：NIC 配置

5. 在 NIC Configuration （NIC 配置）页面上，选择 **Boot Mode** （引导模式），按 ENTER 键，然后选择 **FCoE** 作为首选引导模式。

注

如果在端口级别禁用 **FCoE Mode** （FCoE 模式）功能，则 **FCoE** 不会作为引导选项列出。如果 **Boot Mode** （引导模式）首选为 **FCoE**，请确保 **FCoE Mode** （FCoE 模式）功能已启用，如图 6-29 中所示。并非所有适配器版本均支持 FCoE。

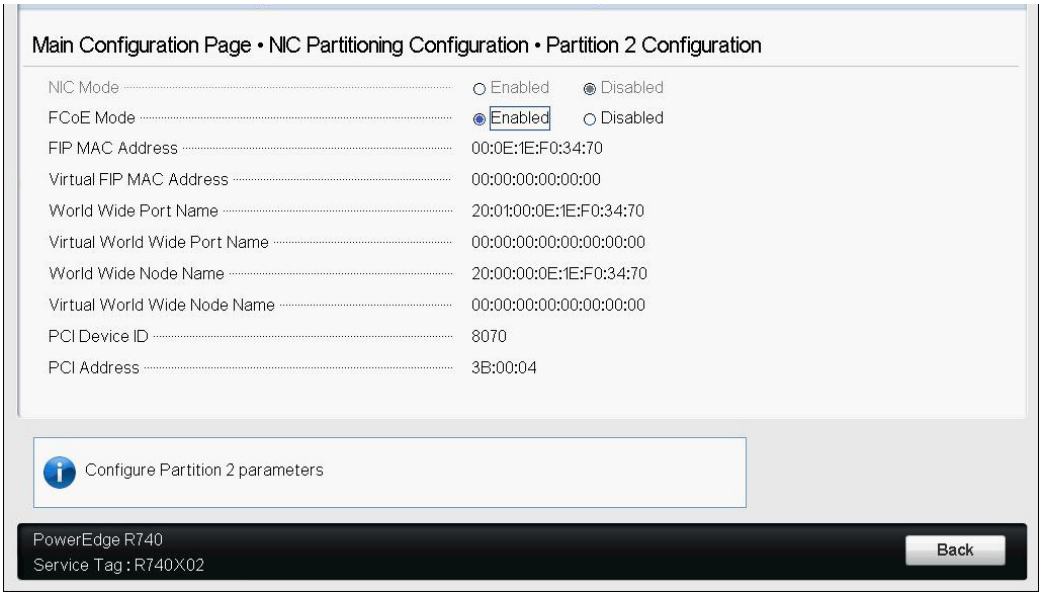


图 6-29. 系统设置：FCoE 模式已启用

要配置 FCoE 引导参数：

1. 在设备 UEFI HII 的 Main Configuration Page （主要配置页面）上，选择 **FCoE Configuration** （FCoE 配置），然后按 ENTER 键。
2. 在 FCoE Configuration （FCoE 配置）页面上，选择 **FCoE General Parameters** （FCoE 常规参数），然后按 ENTER 键。
3. 在 FCoE General Parameters （FCoE 常规参数）页面 （图 6-30）上，按向上箭头键和向下箭头键选择参数，然后按 ENTER 键选择并输入以下值：
   - **Fabric Discovery Retry Count** （结构发现重试计数）：默认值或根据需要
   - **LUN Busy Retry Count** （LUN 忙碌重试计数）：默认值或根据需要

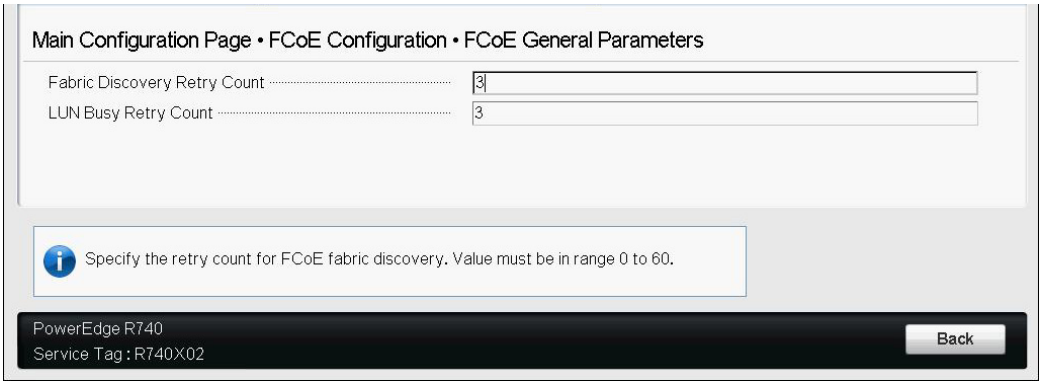


图 6-30. 系统设置：FCoE 常规参数

4. 返回 FCoE Configuration （FCoE 配置）页面。
5. 按 ESC 键，然后选择 **FCoE Target Parameters** （FCoE 目标参数）。
6. 按 ENTER 键。
7. 在 **FCoE General Parameters Menu** （FCoE 常规参数菜单）中，启用到首选 FCoE 目标的 **Connect** （连接）。
8. 键入 FCoE 目标的以下参数的值 （图 6-31），然后按 ENTER 键：
   - ❑ **World Wide Port Name Target *n*** （全局端口名称目标 n）
   - ❑ **Boot LUN *n*** （引导 LUN n）

   其中， *n* 的值介于 1 到 8 之间，使您能够配置 8 个 FCoE 目标。

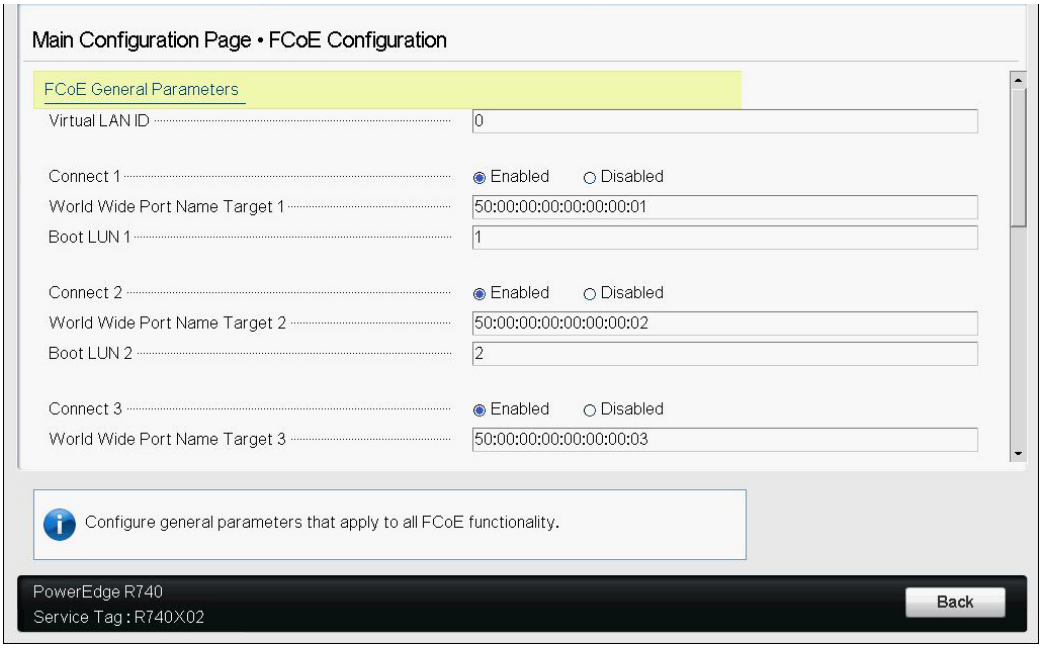


图 6-31. 系统设置：FCoE 常规参数

在 Windows 上配置从 SAN 的 FCoE 引导

Windows 版从 SAN 的 FCoE 引导信息包括：

- Windows Server 2012 R2 和 2016 FCoE 引导安装
- 在 Windows 上配置 FCoE
- Windows 上的 FCoE 故障转储
- 将适配器驱动程序注入 （滑流至） Windows 映像文件中

Windows Server 2012 R2 和 2016 FCoE 引导安装

对于 Windows Server 2012R2/2016 从 SAN 引导安装， Marvell 要求使用 “ 滑流 ” DVD 或 ISO 映像，同时注入最新的 Marvell 驱动程序。请参阅 第 120 页上 “ 将适配器驱动程序注入 （滑流至） Windows 映像文件中 ”。

以下步骤准备在 FCoE 模式下安装和引导的映像。

要设置 Windows Server 2012R2/2016 FCoE 引导：

1. 从要引导的系统 （远程系统）上移除所有本地硬盘驱动器。
2. 通过按照第 120 页上 “ 将适配器驱动程序注入 （滑流至） Windows 映像文件中 ” 中的滑流步骤进行操作，准备 Windows OS 安装介质。
3. 将最新的 Marvell FCoE 引导映像加载到适配器 NVRAM 中。

4. 配置 FCoE 目标以允许从远程设备连接。确保目标有足够磁盘空间安装新的 OS。
5. 配置 UEFI HII 以在所需的适配器端口上设置 FCoE 引导类型、正确的启动器和 FCoE 引导的目标参数。
6. 保存设置并重新引导系统。远程系统应连接至 FCoE 目标，然后从 DVD-ROM 设备引导。
7. 从 DVD 引导并开始安装。
8. 请按照屏幕说明进行操作。

   在显示可用于安装的磁盘列表的窗口上，应会看到 FCoE 目标磁盘。此目标是通过 FCoE 引导协议连接的磁盘，位于远程 FCoE 目标中。
9. 要继续 Windows Server 2012R2/2016 安装，请选择 **Next**（下一步），然后按照屏幕说明进行操作。作为安装过程的一部分，服务器将进行多次重新引导。
10. 您应该在服务器引导至 OS 后，运行驱动程序安装程序以完成 Marvell 驱动程序和应用程序安装。

在 Windows 上配置 FCoE

默认情况下， DCB 在 Marvell FastLinQ 41*xxx* FCoE 和 DCB 兼容 C-NIC 上已启用。 Marvell 41*xxx*FCoE 要求启用 DCB 的接口。对于 Windows 操作系统，使用 QConvergeConsole GUI 或命令行公用程序来配置 DCB 参数。

Windows 上的 FCoE 故障转储

故障转储功能目前支持 FastLinQ 41*xxx* 系列适配器的 FCoE 引导。

在 FCoE 引导模式下， FCoE 故障转储生成无需额外的配置。

将适配器驱动程序注入 （滑流至） Windows 映像文件中

要将适配器驱动程序注入 Windows 映像文件中：

1. 获取适用 Windows Server 版本 （2012、 2012 R2、 2016 或 2019）的最新驱动程序包。
2. 将驱动程序包提取到工作目录：
   a. 打开命令行会话，导航到包含驱动程序包的文件夹。
   b. 要提取 Dell Update Package (DUP) 驱动程序，请发出以下命令：

      ```
      start /wait NameOfDup.exe /s /drivers=<folder path>
      ```
3. 从 Microsoft 下载 Windows 评估和部署工具包 (ADK) 版本 10：

   https://developer.microsoft.com/en-us/windows/hardware/windows-assessment-deployment-kit

4. 遵循 Microsoft 的 “ 向脱机的 Windows 图像添加和删除驱动程序 ” 文章，并注入在 b 部分步骤 2 上提取的 OOB 驱动程序。请参阅 https://docs.microsoft.com/en-us/windows-hardware/manufacture/desktop/add-and-remove-drivers-to-an-offline-windows-image

在 Linux 上配置从 SAN 的 FCoE 引导

Linux 上从 SAN 的 FCoE 引导配置涵盖以下内容：

- Linux FCoE 从 SAN 引导的前提条件
- 配置 Linux FCoE 从 SAN 引导

Linux FCoE 从 SAN 引导的前提条件

使用 Marvell FastLinQ 41*xxx* 10/25GbE 控制器时， Linux FCoE 从 SAN 引导必须满足以下要求才可正确运作。

常规

您不再需要使用 Red Hat 中的 FCoE 磁盘标签和 SUSE 安装程序，因为 FCoE 接口不会显示于网络接口中，而是被 qedf 驱动程序自动激活。

SLES 12 和 SLES 15

- 建议对 SLES 12 SP 3 及更高版本使用驱动程序更新磁盘。
- 为确保安装程序要求驱动程序更新磁盘，必须有安装程序参数 `dud=1`。
- 请勿使用安装程序参数 `withfcoe=1`，因为，如果显示 qede 的网络接口，软件 FCoE 将与硬件卸载冲突。

配置 Linux FCoE 从 SAN 引导

本节为以下 Linux 分发提供从 SAN 的 FCoE 引导步骤：

- 对 SLES 12 SP3 及更高版本配置从 SAN 的 FCoE 引导
- 使用 FCoE 引导设备作为 kdump 目标

对 SLES 12 SP3 及更高版本配置从 SAN 的 FCoE 引导

使用 SLES 12 SP3 时，除了为开箱即用驱动程序注入 DUD 之外，执行从 SAN 安装引导无需其他步骤。

使用 FCoE 引导设备作为 kdump 目标

在使用 qedf 驱动程序用作故障转储的 kdump 目标所显示的设备时， Marvell 建议将内核命令行上的 kdump `crashkernel` 内存参数指定为最小值 512MB。否则内核故障转储可能不成功。

有关如何设置 kdump `crashkernel` 大小的详细信息，请参阅 Linux 分发文档。

在 VMware 上配置 SAN 的 FCoE 引导

对于 VMware ESXi 6.5/6.7 从 SAN 引导安装， Marvell 要求您使用注入最新 Marvell 聚合网络适配器驱动程序包而创建的自定义的 ESXi ISO 映像。本节涵盖 VMware FCoE 从 SAN 引导的以下步骤。

- 将 (滑溜至) ESXi 适配器驱动程序注入到映像文件
- 安装自定义的 ESXi ISO

将 (滑溜至) ESXi 适配器驱动程序注入到映像文件

此步骤以 ESXi-Customizer 工具 v2.7.2 为例，但您可以使用任何 ESXi 自定义程序。

要将适配器驱动程序注入 ESXi 映像文件中：

1. 下载 ESXi-Customizer v2.7.2 或更高版本。
2. 转到 `ESXi customizer` 目录。
3. 发出 **`ESXi-Customizer.cmd`** 命令。
4. 在 ESXi-Customizer 对话框中，单击 **Browse** （浏览）完成以下步骤。
   a. 选择原始 VMware ESXi ISO 文件。
   b. 选择 Marvell FCoE 驱动程序 VIB 文件或 Marvell 脱机 qedentv 包 ZIP 文件。
   c. 对于工作目录，选择需要在其中创建自定义的 ISO 的文件夹。
   d. 单击 **Run** （运行）。

图 6-32 显示一个示例。

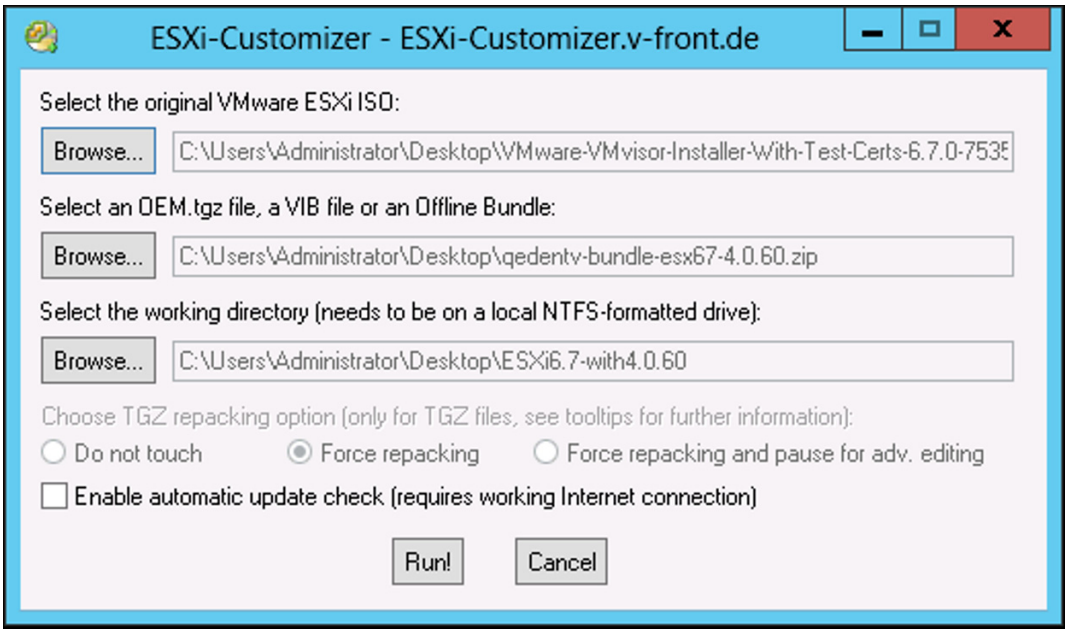


图 6-32. ESXi-Customizer 对话框

5. 刻录包含位于步骤 4c 指定的工作目录中的自定义的 ISO 版本的 DVD。
6. 使用新的 DVD 安装 ESXi OS。

安装自定义的 ESXi ISO

1. 将最新的 Marvell FCoE 引导映像加载到适配器 NVRAM 中。
2. 配置 FCOE 目标以允许有效地连接远程计算机。确保目标有足够的可用磁盘空间安装新的 OS。
3. 配置 UEFI HII 以在所需的适配器端口上设置 FCOE 引导类型、正确的启动器和 FCOE 引导的目标参数。
4. 保存设置并重新引导系统。

   启动器应连接至 FCOE 目标，然后从 DVD-ROM 设备引导系统。
5. 从 DVD 引导并开始安装。
6. 请按照屏幕说明进行操作。

在显示可用于安装的磁盘列表的窗口上， FCOE 目标磁盘应该可见，因为注入的聚合网络适配器驱动程序包在自定义的 ESXi ISO 中。图 6-33 显示一个示例。

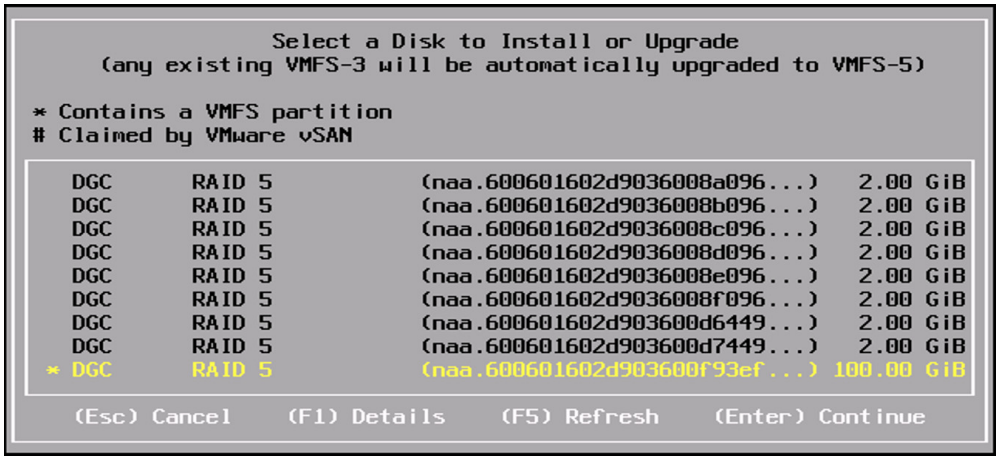
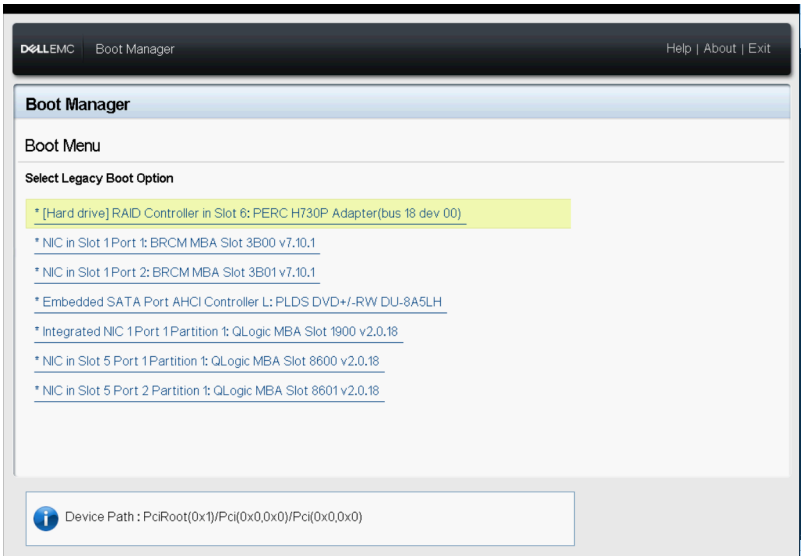


图 6-33. 选择要安装的 VMware 磁盘

7. 选择 ESXi 可在其中安装的 LUN，然后按 ENTER 键。
8. 在下一个窗口上，单击 **Next** （下一步），然后按照屏幕上的说明操作。
9. 当安装完成时，重新引导服务器并弹出 DVD。
10. 在服务器引导时，按 F9 键访问 **One-Time Boot Menu** （一次性引导菜单），然后选择 **Boot media to QLogic adapter port** （引导介质到 Qlogic 适配器端口）。
11. 在 **Boot Menu** （引导菜单）下，选择新安装的 ESXi 以通过从 SAN 引导加载。

图 6-34 提供了两个示例。



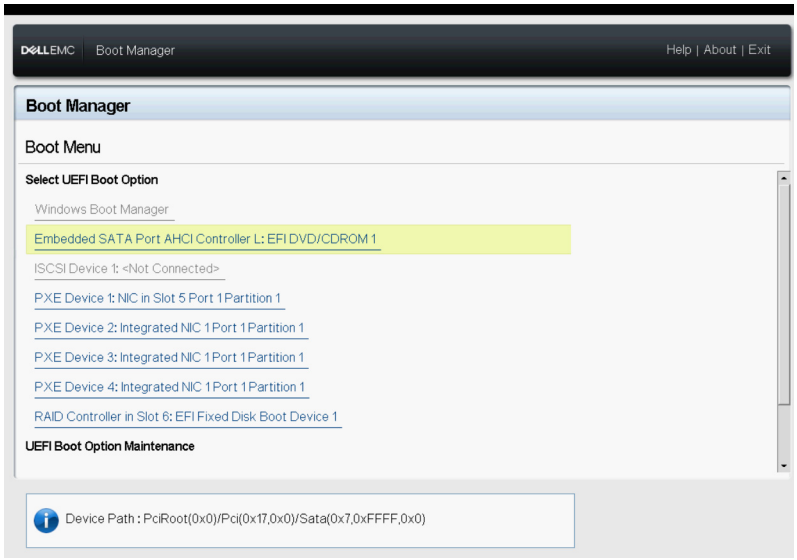


图 6-34. VMware USB 引导选项

7 RoCE 配置

本章介绍 41*xxx* 系列适配器、以太网交换机以及 Windows、Linux 或 VVMware 主机上基于聚合以太网的 RDMA （RoCE v1 和 v2）配置，包括以下内容：

注

某些 RoCE 功能在当前版本中可能并未完全启用。

支持的操作系统和 OFED

表 7-1 显示 RoCE v1、RoCE v2、iWARP 和 OpenFabrics 企业分布（OFED）的操作系统支持。 Windows 或 VMware ESXi 不支持 OFED。

表 7-1. RoCE v1、 RoCE v2、 iWARP 、 iSER 和 OFED 的 OS 支持

操作系统	内建	OFED-4.17-1 GA
Windows Server 2012	N/A	N/A
Windows Server 2012 R2	否	N/A
Windows Server 2016	否	N/A
Windows Server 2019	RoCE v1、 RoCE v2、 iWARP	N/A
RHEL 7.6	RoCE v1、RoCE v2、iWARP、iSER	RoCE v1、 RoCE v2、 iWARP

表 7-1. RoCE v1、 RoCE v2、 iWARP 、 iSER 和 OFED 的 OS 支持 (续)

操作系统	内建	OFED-4.17-1 GA
RHEL 7.7	RoCE v1、RoCE v2、iWARP、iSER	否
RHEL 8.0	RoCE v1、RoCE v2、iWARP、iSER	否
RHEL 8.1	RoCE v1、RoCE v2、iWARP、iSER	否
SLES 12 SP4	RoCE v1、RoCE v2、iWARP、iSER	RoCE v1、 RoCE v2、 iWARP
SLES 15 SP0	RoCE v1、RoCE v2、iWARP、iSER	RoCE v1、 RoCE v2、 iWARP
SLES 15 SP1	RoCE v1、RoCE v2、iWARP、iSER	否
CentOS 7.6	RoCE v1、RoCE v2、iWARP、iSER	RoCE v1、 RoCE v2、 iWARP
VMware ESXi 6.5 U3	RoCE v1、 RoCE v2	N/A
VMware ESXi 6.7 U2	RoCE v1、 RoCE v2	N/A

计划 RoCE

在准备执行 RoCE 时，请考虑以下限制：

- 如果使用的是内建 OFED，则服务器和客户端系统上的操作系统应该相同。一些应用程序也许可以在不同的操作系统之间正常工作，但无法保证。这是 OFED 限制。
- 对于 OFED 应用程序 （最常见的是 perftest 应用程序），服务器和客户端应用程序应使用相同的选项和值。如果操作系统和 perftest 应用程序具有不同的版本，则可能会出现问题。要确认 perftest 版本，请发出以下命令：

  ```
  # ib_send_bw --version
  ```

- 在内建 OFED 中生成 libqedr 需要安装 libibverbs-devel。
- 在内建 OFED 中运行用户空间应用程序需要安装 InfiniBand® 支持组，通过 yum groupinstall， “InfiniBand 支持 ” 包含 libibcm、 libibverbs 等等。
- 依赖于 libibverbs 的 OFED 和 RDMA 应用程序也需要 Marvell RDMA 用户空间库 libqedr。使用 libqedr RPM 或源文件包安装 libqedr。
- RoCE 仅支持小字节序。

准备适配器

按照以下步骤操作以启用 DCBX，并使用 HII 管理应用程序指定 RoCE 优先级。有关 HII 应用程序的信息，请参阅第 5 章 适配器预引导配置。

准备适配器：

1. 在 Main Configuration Page （主要配置页面）上，选择 **Data Center Bridging (DCB) Settings** （数据中心桥接 (DCB) 设置），然后单击 **Finish** （完成）。
2. 在 Data Center Bridging (DCB) Settings （数据中心桥接 (DCB) 设置）窗口中，单击 **DCBX Protocol**（DCBX 协议）选项。41*xxx* 系列适配器同时支持 CEE 和 IEEE 协议。此值应该符合 DCB 交换机上的相应值。在本示例中，选择 **CEE** 或 **Dynamic** （动态）。
3. 在 **RoCE Priority** （RoCE 优先级）框中，键入优先级值。此值应该符合 DCB 交换机上的相应值。在本示例中，键入 **5**。通常， 0 用于默认有损流量类， 3 用于 FCoE 流量类， 4 用于通过 DCB 的无损 iSCSI-TLV 流量类。
4. 单击 **Back** （后退）。
5. 看到提示时，单击 **Yes** （是）以保存更改。更改将在系统重设后生效。

   在 Windows 中，可使用 HII 或 QoS 方法配置 DCBX。本节所示配置通过 HII 实现。对于 QoS，请参阅第 256 页上 “ 为 RoCE 配置 QoS”。

准备以太网交换机

本节介绍如何为 RoCE 配置 Cisco® Nexus® 6000 以太网交换机和 Dell Z9100 以太网交换机。

- 配置 Cisco Nexus 6000 以太网交换机
- 为 RoCE 配置 Dell Z9100 以太网交换机

配置 Cisco Nexus 6000 以太网交换机

为 RoCE 配置 Cisco Nexus 6000 以太网交换机步骤涉及配置类映射、配置策略映射、应用策略，以及为交换机端口分配 vLAN ID。

要配置 Cisco 交换机：

1. 执行以下步骤打开配置终端会话：

```
Switch# config terminal
switch(config)#
```

2. 执行以下步骤配置相关服务质量 (QoS) 类映射，并设置 RoCE 优先级 (cos) 以匹配适配器 (5)：

```
switch(config)# class-map type qos class-roce
switch(config)# match cos 5
```

3. 执行以下步骤配置排队类映射：

```
switch(config)# class-map type queuing class-roce
switch(config)# match qos-group 3
```

4. 执行以下步骤配置网络 QoS 类映射：

```
switch(config)# class-map type network-qos class-roce
switch(config)# match qos-group 3
```

5. 执行以下步骤配置 QoS 策略映射：

```
switch(config)# policy-map type qos roce
switch(config)# class type qos class-roce
switch(config)# set qos-group 3
```

6. 配置排队策略映射以分配网络带宽。在本示例中，使用的值为 50%：

```
switch(config)# policy-map type queuing roce
switch(config)# class type queuing class-roce
switch(config)# bandwidth percent 50
```

7. 执行以下步骤配置网络 QoS 策略映射以设置无丢弃流量类的优先级流控制：

```
switch(config)# policy-map type network-qos roce
switch(config)# class type network-qos class-roce
switch(config)# pause no-drop
```

8. 执行以下步骤在系统级别应用新策略：

```
switch(config)# system qos
switch(config)# service-policy type qos input roce
switch(config)# service-policy type queuing output roce
switch(config)# service-policy type queuing input roce
switch(config)# service-policy type network-qos roce
```

9. 为交换机端口分配 vLAN ID，以匹配分配给适配器 (5) 的 vLAN ID。

```
switch(config)# interface ethernet x/x
switch(config)# switchport mode trunk
switch(config)# switchport trunk allowed vlan 1,5
```

为 RoCE 配置 Dell Z9100 以太网交换机

为 RoCE 配置 Dell Z9100 以太网交换机涉及为 RoCE 配置 DCB 映射、配置基于优先级的流控制 (PFC) 和增强的传输选择 (ETS)、验证 DCB 映射、将 DCB 映射应用至端口、验证端口上的 PFC 和 ETS、指定 DCB 协议，以及为交换机端口分配 VLAN ID。

注

有关配置 Dell Z9100 交换机端口以便按 25 Gbps 速率连接 41xxx 系列适配器的说明，请参见第 294 页上 “Dell Z9100 交换机配置 ”。

要配置 Dell 交换机：

1. 创建 DCB 映射。

```
Dell# configure
Dell(conf)# dcb-map roce
Dell(conf-dcbmap-roce)#
```

2. 在 DCB 映射中配置两个 ETS 流量类，并为 RoCE （组 1）分配 50% 带宽。

```
Dell(conf-dcbmap-roce)# priority-group 0 bandwidth 50 pfc off
Dell(conf-dcbmap-roce)# priority-group 1 bandwidth 50 pfc on
```

3. 配置 PFC 优先级以符合适配器流量等级优先级编号 (5)。

```
Dell(conf-dcbmap-roce)# priority-pgid 0 0 0 0 0 1 0 0
```

4. 验证 DCB 映射配置优先级组。

```
Dell(conf-dcbmap-roce)# do show qos dcb-map roce
-----------------------
State   :Complete
PfcMode :ON
--------------------
PG:0 TSA:ETS  BW:40  PFC:OFF
Priorities:0 1 2 3 4 6 7

PG:1 TSA:ETS  BW:60  PFC:ON
Priorities:5
```

5. 将 DCB 映射应用到端口。

```
Dell(conf)# interface twentyFiveGigE 1/8/1
Dell(conf-if-tf-1/8/1)# dcb-map roce
```

6. 验证端口上的 ETS 和 PFC 配置。以下示例显示了 ETS 的汇总接口信息和 PFC 的详细接口信息

```
Dell(conf-if-tf-1/8/1)# do show interfaces twentyFiveGigE 1/8/1 ets summary
Interface twentyFiveGigE 1/8/1
Max Supported TC is 4
Number of Traffic Classes is 8
Admin mode is on

Admin Parameters :
------------------
Admin is enabled

PG-grp  Priority#            BW-%      BW-COMMITTED       BW-PEAK  TSA
            %            Rate(Mbps) Burst(KB) Rate(Mbps) Burst(KB)
-------------------------------------------------------------------------
0       0,1,2,3,4,6,7    40   -        -         -            -    ETS
1       5                60   -        -         -            -    ETS
2                        -    -        -         -            -     -
3                        -    -        -         -            -     -

Dell(Conf)# do show interfaces twentyFiveGigE 1/8/1 pfc detail
Interface twentyFiveGigE 1/8/1
     Admin mode is on
     Admin is enabled, Priority list is 5
     Remote is enabled, Priority list is 5
     Remote Willing Status is enabled
     Local is enabled, Priority list is 5
     Oper status is init
     PFC DCBX Oper status is Up
     State Machine Type is Feature
     TLV Tx Status is enabled
     PFC Link Delay 65535 pause quntams
     Application Priority TLV  Parameters :
     -------------------------------------
     FCOE TLV Tx Status is disabled
     ISCSI TLV Tx Status is enabled
     Local FCOE PriorityMap is 0x0
     Local ISCSI PriorityMap is 0x20
     Remote ISCSI PriorityMap is 0x200
```

```
66 Input TLV pkts, 99 Output TLV pkts, 0 Error pkts, 0 Pause Tx pkts, 0 Pause
Rx pkts
66 Input Appln Priority TLV pkts, 99 Output Appln Priority TLV pkts, 0 Error
Appln Priority TLV Pkts
```

7. 配置 DCBX 协议 （本例中为 CEE）。

```
Dell(conf)# interface twentyFiveGigE 1/8/1
Dell(conf-if-tf-1/8/1)# protocol lldp
Dell(conf-if-tf-1/8/1-lldp)# dcbx version cee
```

8. 为交换机端口分配 VLAN ID，以匹配分配给适配器 (5) 的 VLAN ID。

```
Dell(conf)# interface vlan 5
Dell(conf-if-vl-5)# tagged twentyFiveGigE 1/8/1
```

注

VLAN ID 不必与 RoCE 流量等级优先级编号相同。然而，使用相同的数字会使配置更容易理解。

在 Windows Server 的适配器上配置 RoCE

在 Windows Server 主机的适配器上配置 RoCE 涉及在适配器上启用 RoCE 以及验证 Network Direct MTU 大小。

要在 Windows Server 主机上配置 RoCE：

1. 在适配器上启用 RoCE。
   a. 打开 Windows 设备管理器，然后打开 41*xxx* 系列适配器 NDIS 微型端口属性。
   b. 在 QLogic FastLinQ Adapter Properties （QLogic FastLinQ 适配器属性）上，单击 **Advanced** （高级）选项卡。
   c. 在 Advanced （高级）页面上，通过选择 **Property** （属性）下的每个项目，然后选择该项目相应的 **Value** （值），配置表 7-2 中列出的属性。然后单击 **OK** （确定）。

表 7-2. RoCE 的高级属性

属性	值或说明
NetworkDirect Functionality （NetworkDirect 功能）	**Enabled** （已启用）

表 7-2. RoCE 的高级属性 (续)

属性	值或说明
Network Direct Mtu Size（Network Direct Mtu 大小）	network direct MTU 大小必须小于巨型数据包大小。
Quality of Service（相关服务质量）	配置 RoCE v1/v2 时，始终选择 **Enabled**（已启用）允许 Windows DCB-QoS 服务控制和监测 DCB。有关更多信息，请参阅第 257 页上 “ 通过在适配器上禁用 DCBX 配置 QoS” 和第 261 页上 “ 通过在适配器上启用 DCBX 配置 QoS”。
NetworkDirect Technology（NetworkDirect 技术）	**RoCE** 或 **RoCE v2**。
VLAN ID	将任何 vLAN ID 分配给接口。该值必须与交换机上分配的值相同。

图 7-1 显示配置属性值的示例。

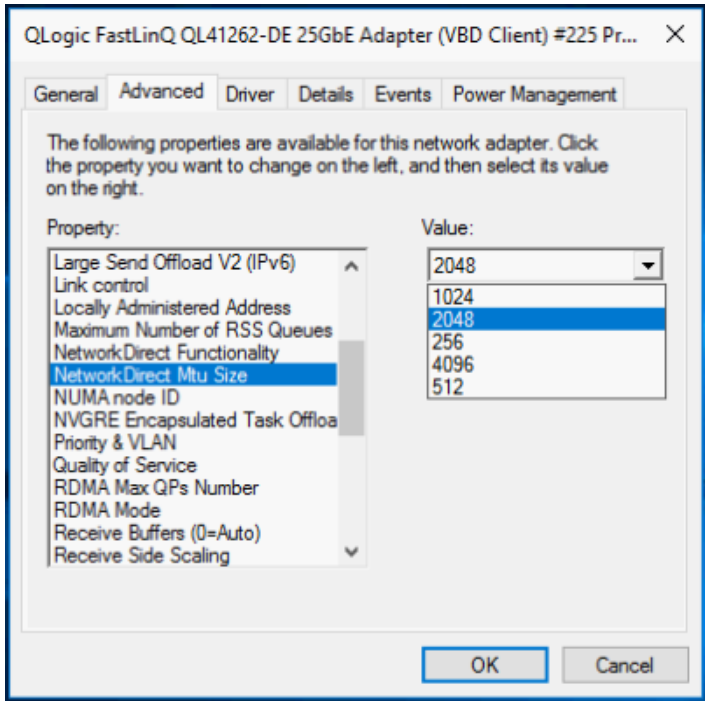


图 7-1. 配置 RoCE 属性

2. 使用 Windows PowerShell，验证适配器上已启用 RDMA。
Get-NetAdapterRdma 命令列出支持 RDMA 的适配器 - 两个端口同时启用。

注

如果通过 Hyper-V 配置 RoCE，请勿将 vLAN ID 分配给物理接口。

```
PS C:\Users\Administrator> Get-NetAdapterRdma
Name                InterfaceDescription         Enabled
-----  ------  -----------  --------------
SLOT 4 3 Port 1    QLogic FastLinQ QL41262...   True
SLOT 4 3 Port 2    QLogic FastLinQ QL41262...   True
```

3. 使用 Windows PowerShell 验证在主机操作系统上是否启用了 NetworkDirect。 Get-NetOffloadGlobalSetting 命令显示 NetworkDirect 已启用。

```
PS C:\Users\Administrators> Get-NetOffloadGlobalSetting
ReceiveSideScaling            : Enabled
ReceiveSegmentCoalescing      : Enabled
Chimney                       : Disabled
TaskOffload                   : Enabled
NetworkDirect                 : Enabled
NetworkDirectAcrossIPSubnets  : Blocked
PacketCoalescingFilter        : Disabled
```

4. 连接服务器消息块 (SMB) 驱动器，运行 RoCE 流量并验证结果。

   要设置并连接到 SMB 驱动器，请查看 Microsoft 在线提供的信息：

   https://technet.microsoft.com/en-us/library/hh831795(v=ws.11).aspx

5. 默认情况下， Microsoft 的 SMB Direct 为每个端口建立两个 RDMA 连接，以提供优质性能，包括较高块大小 （例如 64KB）下的线速率。为优化性能，您可以将每个 RDMA 接口的 RDMA 连接数更改为四个 （或更多）。

   要将 RDMA 连接数增加至四个 （或更多），请在 Windows PowerShell 中发出以下命令：

```
PS C:\Users\Administrator> Set-ItemProperty -Path
"HKLM:\SYSTEM\CurrentControlSet\Services\LanmanWorkstation\
Parameters" ConnectionCountPerRdmaNetworkInterface -Type
DWORD -Value 4 -Force
```

查看 RDMA 计数器

以下步骤也适用于 iWARP。

为了查看 RoCE 的 RDMA 计数器：

1. 启动性能监测器。
2. 打开 Add Counters （添加计数器）对话框。图 7-2 显示一个示例。

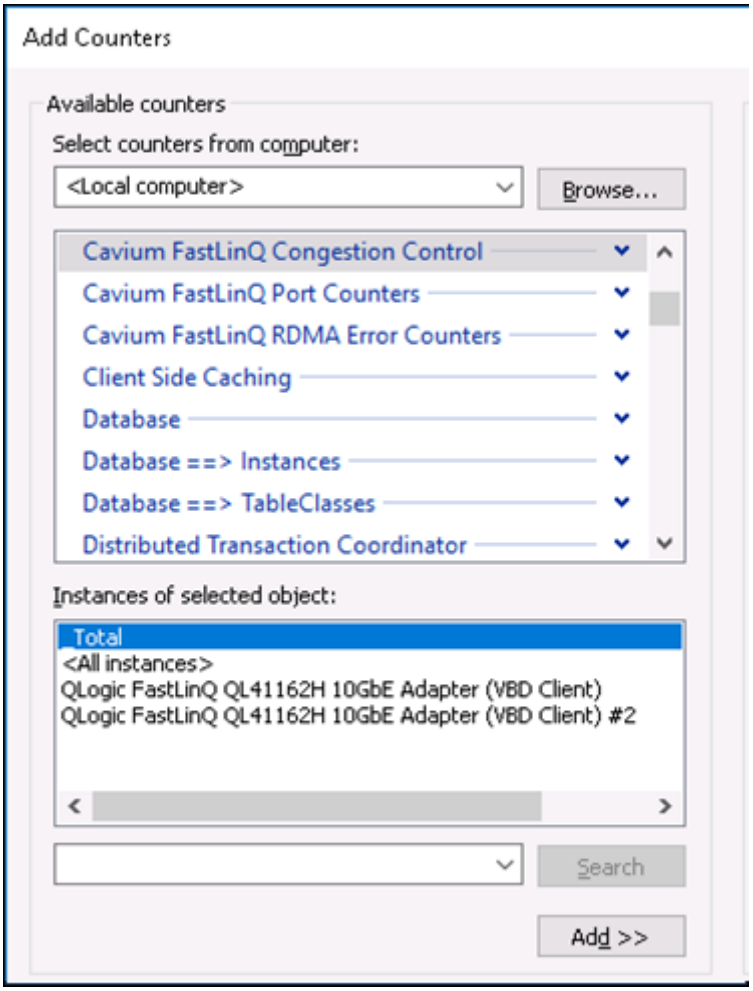


图 7-2. Add Counters （添加计数器）对话框

注

如果 Marvell RDMA 计数器没有列在 Performance Monitor （性能监视器） Add Counters （添加计数器）对话框中，请通过从驱动程序位置发出以下命令手动添加：

```
Lodctr /M:qend.man
```

3. 选择以下计数器类型之一：

   - **Cavium FastLinQ Congestion Control** （拥塞控制）：
     - 当网络中出现拥塞并且交换机上启用 ECN 时的增量。
     - 描述成功发送和接收的 RoCE v2 ECN 标记数据包和拥塞通知数据包 (CNP)。
     - 仅适用于 RoCE v2。
   - **Cavium FastLinQ Port Counters** （端口计数器）：
     - 当网络出现拥塞时的增量。
     - 当配置流量控制或全局暂停且网络出现拥塞时，暂停计数器。
     - 当配置优先级流量控制且网络出现拥塞时的 PFC 计数器增量。
   - **Cavium FastLinQ RDMA Error Counters** （错误计数器）：
     - 传输操作出现任何错误时的增量。
     - 有关详细信息，请参阅 表 7-3。

4. 在 **Instances of selected object** （所选对象的实例）下，选择 **Total** （总计），然后单击 **Add** （添加）。

图 7-3 显示了 计数器监测输出的三个示例。

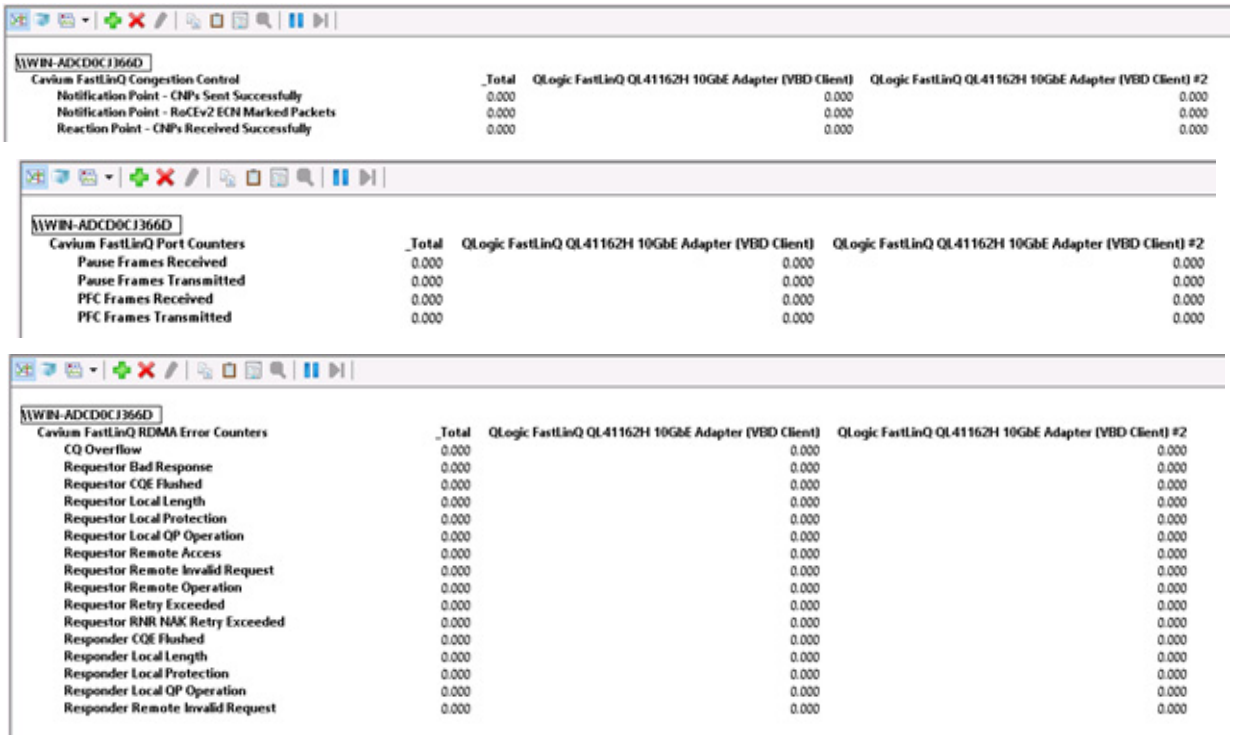


图 7-3. 性能监视：41xxx 系列适配器计数器

表 7-3 提供错误计数器的有关详细信息。

表 7-3. Marvell FastLinQ RDMA 错误计数器

RDMA 错误计数器	说明	适用于 RoCE?	适用于 iWARP?	故障排除
CQ overflow（CQ 溢出）	发布 RDMA 工作请求的完成队列。该计数器指定在发送或接收队列上的工作请求已完成，但在关联的完成队列中没有空间的实例数量。	是	是	表明导致完成队列空间不足的软件设计问题。
Requestor Bad response（请求者响应差）	响应者返回故障响应。	是	是	—

表 7-3. Marvell FastLinQ RDMA 错误计数器 (续)

RDMA 错误计数器	说明	适用于 RoCE?	适用于 iWARP?	故障排除
`Requestor CQEs flushed with error`（请求者 CQE 发生错误）	当 QP 因任何原因进入错误状态并且存在待处理的工作请求时，可通过向 CQ 发送处于刷新状态的完成 （而非完成实际的工作请求）来刷新发布的工作请求。如果工作请求在错误状态下完成，则该 QP 的所有其它待处理工作请求将刷新。	是	是	关闭 RDMA 连接时出现。
`Requestor Local length`（请求者的本地长度）	包含太多或太少有效载荷数据的 RDMA Read 响应消息。	是	是	通常表明主机软件组件相关的问题。
`Requestor local protection`（请求者的本地保护）	本地发布的工作请求数据段不引用对所请求操作有效的内存区域。	是	是	通常表明主机软件组件相关的问题。
`Requestor local QP operation`（请求者本地 QP 操作）	处理此工作请求时检测到内部 QP 一致性错误。	是	是	—
`Requestor Remote access`（请求者的远程访问）	远程数据缓冲区出现保护错误，该错误可通过 RDMA Read 读取、通过 RDMA Write 写入或通过原子操作访问。	是	是	—
`Requestor Remote Invalid request`（请求者的远程无效请求）	远程端收到信道上的无效消息。无效请求可为 Send （发送）消息或 RDMA 请求。	是	是	可能的原因包括：该接收队列不支持此操作；没有足够的缓冲来接收新的 RDMA 或原子操作请求；或 RDMA 请求所指定的长度大于 231 字节。
`Requestor remote operation`（请求者的远程操作）	由于其本地问题，远程端无法完成所请求的操作。	是	是	出现在远程端的阻止操作完成的软件问题（例如，在 RQ 上引发 QP 错误或 WQE 故障的软件问题）。

表 7-3. Marvell FastLinQ RDMA 错误计数器 (续)

RDMA 错误计数器	说明	适用于 RoCE?	适用于 iWARP?	故障排除
Requestor retry exceeded（请求者重试次数超限）	传输重试次数已超过最大限制。	是	是	远程对等方可能已停止响应，或网络问题正阻止消息确认。
Requestor RNR Retries exceeded（请求者 RNR 重试次数超限）	接收到的 RNR NAK 重试次数已达上限，且无一次成功。	是	否	远程对等方可能已停止响应，或网络问题正阻止消息确认。
Responder CQE flushed（响应器 CQE 已刷新）	当 QP 因任何原因进入错误状态并且 RQ 上存在待处理接收缓存区时，可通过向 CQ 发送处于刷新状态的完成来刷新发布的工作请求 （RQ 上的接收缓冲区）。如果工作请求在错误状态下完成，那么该 QP 的所有其它待处理的工作请求将刷新。	是	是	—
Responder local length（响应器的本地长度）	入站消息中的无效长度。	是	是	恶意的远程对等机。例如，入站发送消息的长度大于接收缓冲区的大小。
Responder local protection（响应器的本地保护）	本地发布的工作请求数据段不引用对所请求操作有效的内存区域。	是	是	表明内存管理相关的软件问题。
Responder Local QP Operation error（响应器本地 QP 操作错误）	处理此工作请求时检测到内部 QP 一致性错误。	是	是	表明软件问题。
Responder remote invalid request（响应器远程无效请求）	响应器在信道上检测到无效入站消息。	是	是	提示来自远程对等机的可能的恶意行为。可能的原因包括：该接收队列不支持此操作；没有足够的缓冲来接收新的 RDMA 请求；或 RDMA 请求所指定的长度大于 2^{31} 字节。

为 SR-IOV VF 设备 (VF RDMA) 配置 RoCE

以下章节介绍如何为 SR-IOV VF 设备 （也称作 *VF RDMA*）配置 RoCE。同时也提供相关信息和限制。

配置说明

要配置 VF RDMA：

1. 安装 VF RDMA 功能组件 （驱动程序、固件、多引导映像 (MBI)）。
2. 为 VF RDMA 配置 QoS。

   需 QoS 配置来配置 RDMA 的优先级流控制 (PFC)。如第 256 页上 “ 为 RoCE 配置 QoS” 中所述，在主机上配置 QoS。（QoS 配置必须在主机完成，而不是在 VM 中进行。）
3. 为 VF RDMA 配置 Windows Hyper-V：
   a. 在 HII 以及 Windows Device Manager （Windows 设备管理器）中的 Advanced （高级）选项卡上启用 SR-IOV。
   b. 打开主机上的 **Windows Hyper-V Manager** （Windows Hyper-V 管理器）。
   c. 在右侧窗格中打开 **Virtual Switch Manager** （虚拟交换机管理器）。

d. 选择 **New Virtual Network switch**（新虚拟网络交换机），并选择 **External**（外部）类型。

图 7-4 显示一个示例。

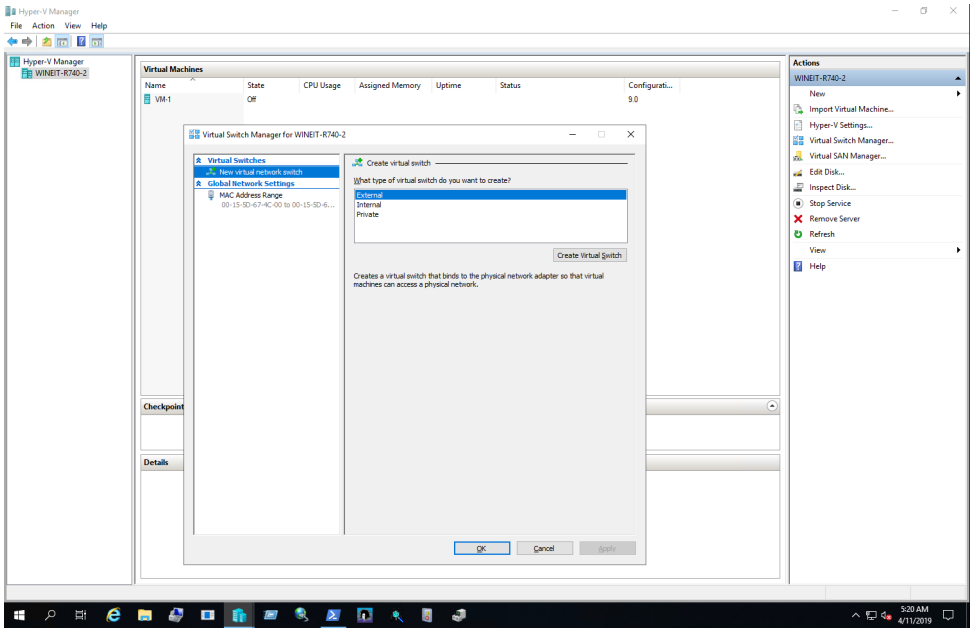


图 7-4. 设置外部新虚拟网络交换机

e. 单击 **External network**（外部网络）按钮，然后选择相应适配器。单击 **Enable single-root I/O virtualization**（启用单根 I/O 虚拟化）**(SR-IOV)**。

图 7-5 显示一个示例。

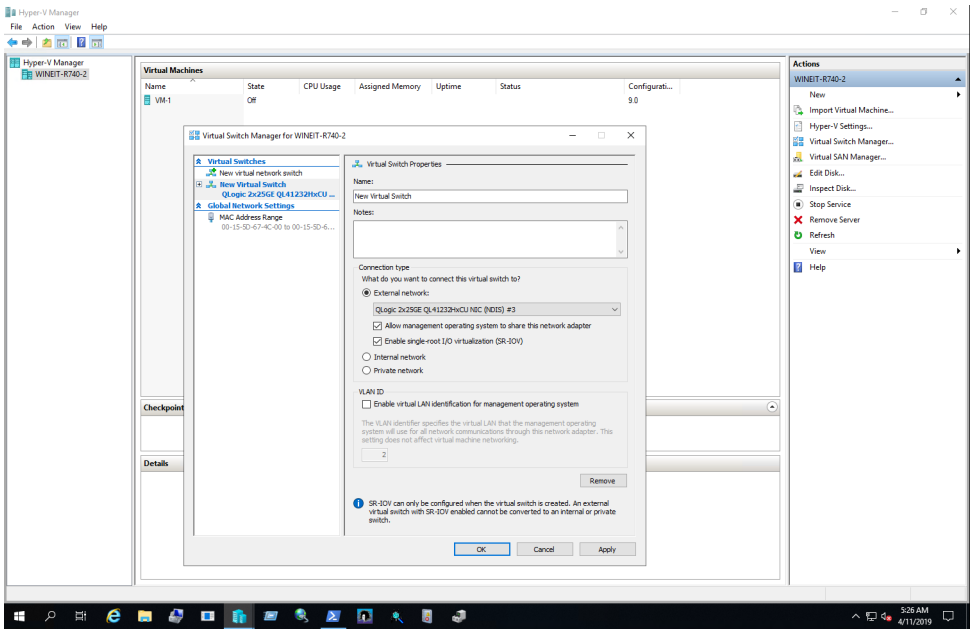


图 7-5. 为新虚拟交换机设置 SR-IOV

f. 创建 VM 并打开 VM 设置。

图 7-6 显示一个示例。

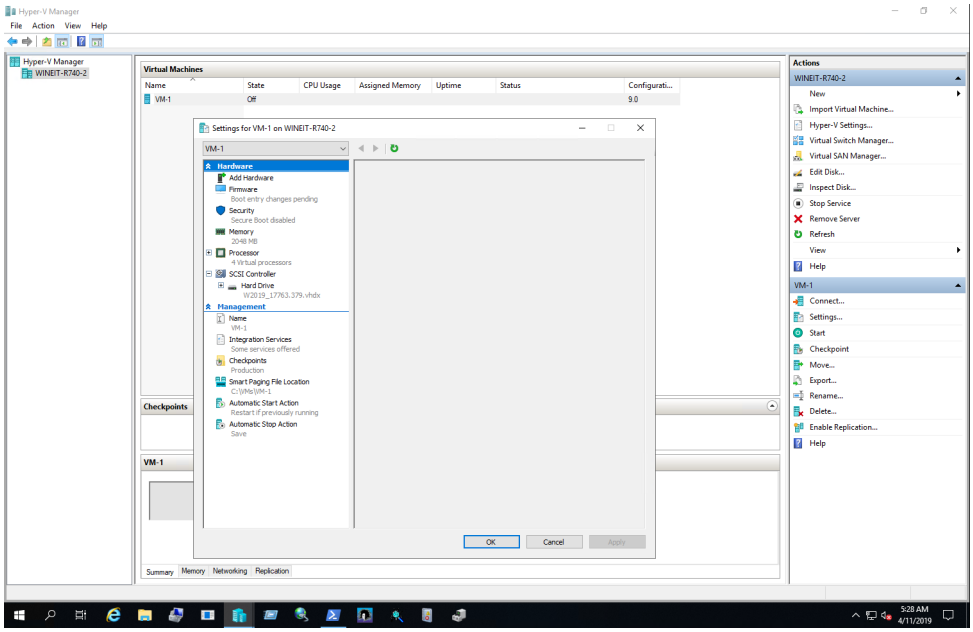


图 7-6. VM 设置

g. 选择 **Add Hardware**（添加硬件），然后选择 **Network Adapter**（网络适配器）分配虚拟网络适配器 (VMNIC) 至 VM。

h. 选择新创建的虚拟交换机。

i. 启用网络适配器的 VLAN。

图 7-7 显示一个示例。

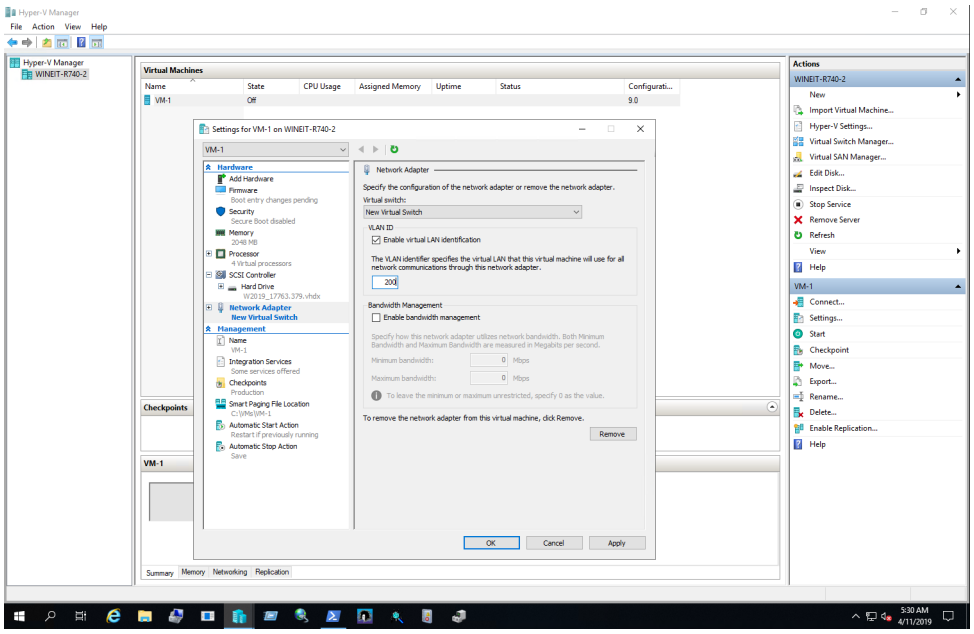


图 7-7. 启用网络适配器的 VLAN

j. 扩展网络适配器设置。在 Single-root I/O virtualization （单根 I/O 虚拟化）下，选择 **Enable SR-IOV** （启用 SR-IOV）启用 VMNIC 的 SR-IOV 功能。

图 7-8 显示一个示例。

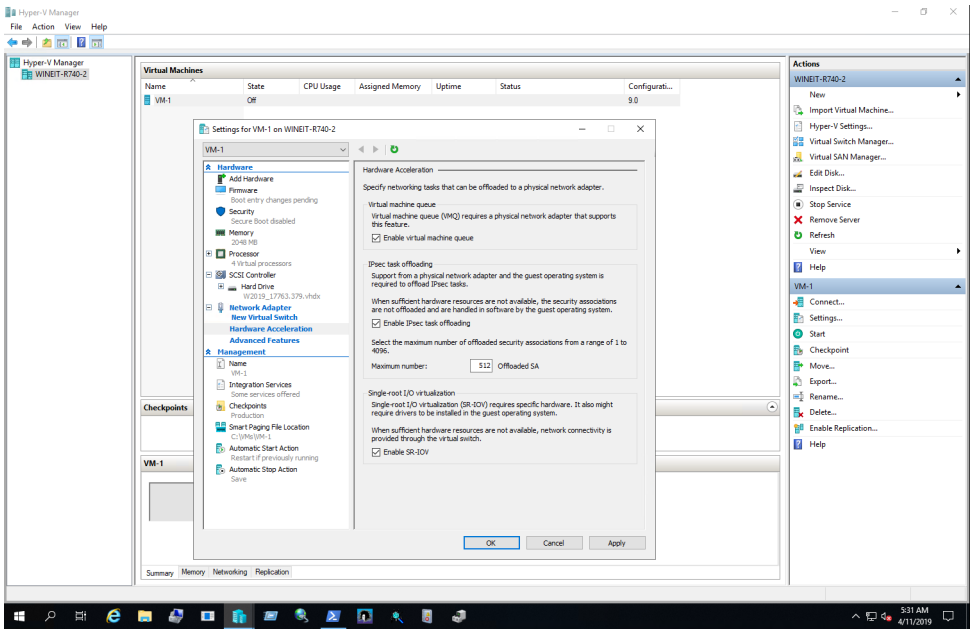


图 7-8. 启用网络适配器的 SR-IOV

4. 在主机上发出以下 PowerShell 命令启用 VMNIC (SR-IOV VF) 的 RDMA 功能。

```
Set-VMNetworkAdapterRdma -VMName <VM_NAME>
-VMNetworkAdapterName <VM_NIC_NAME> -RdmaWeight 100
```

注

发出 PowerShell 命令前必须关闭 VM。

5. 通过引导 VM 和使用 Marvell CD 上的 Windows Super Installer （Windows 超级安装程序）安装最新驱动程序以升级 VM 中的 Marvell 驱动程序。

   图 7-9 显示一个示例。

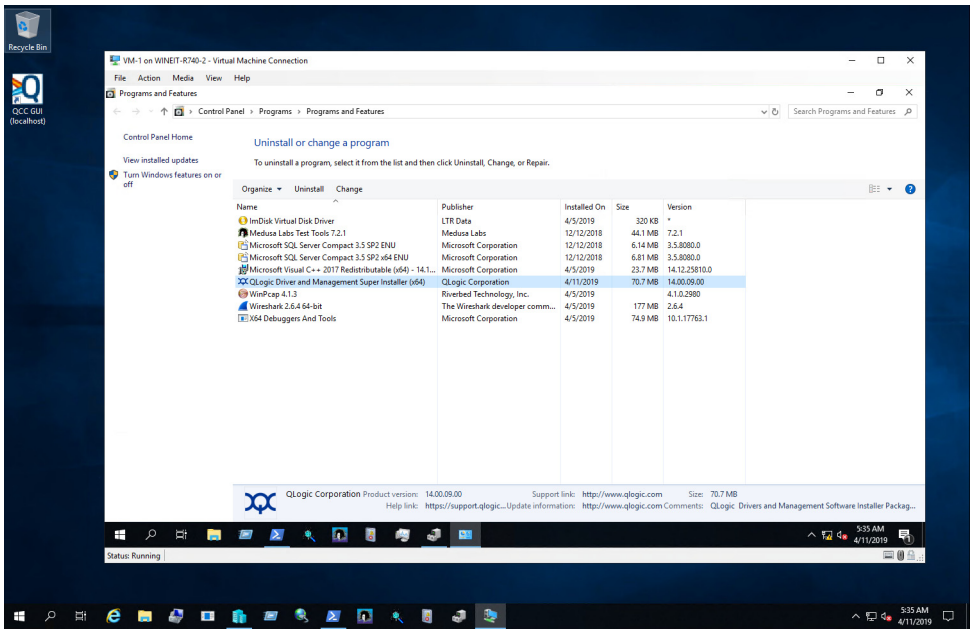


图 7-9. 升级 VM 中的驱动程序

6. 启用与 VM 内部的 VF 关联的 Microsoft 网络设备上的 RMDA。

   图 7-10 显示一个示例。

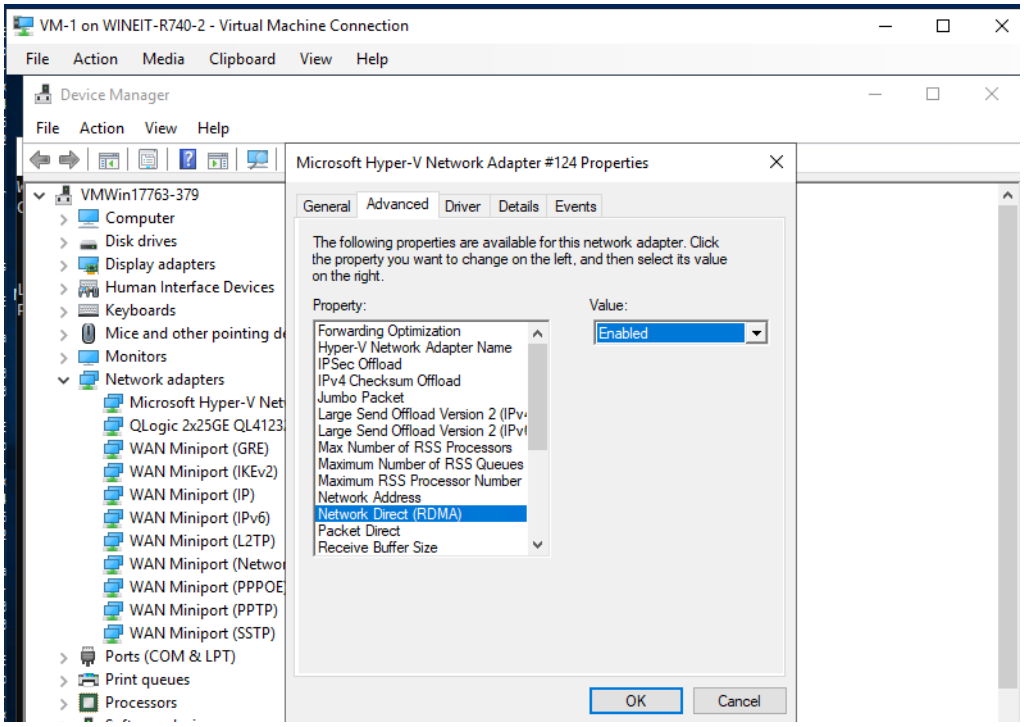


图 7-10. 在 VMNIC 上启用 RDMA

7. 启动 VM RMDA 流量：

   a. 连接服务器消息块 (SMB) 驱动器，运行 RoCE 流量并验证结果。

   b. 打开 VM 中的性能监测器，然后添加 **RDMA Activity counter**（RDMA 活动计数器）。

c. 验证 RDMA 流量正在运行。

图 7-11 提供一个示例。

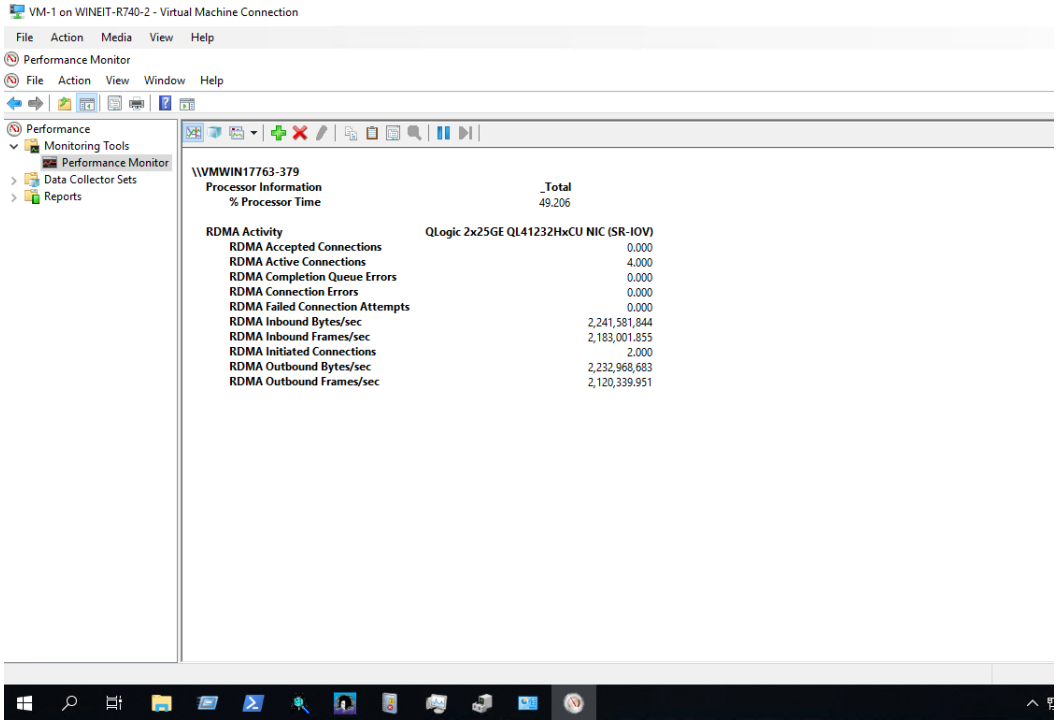


图 7-11. RDMA 流量

限制

VF RDMA 有以下限制：

- 仅基于 41*xxx* 的设备支持 VF RDMA。
- 在当前发布阶段，仅 RoCEv2 支持 VF RDMA。必须对主机和 VM 中的 SR-IOV VF 的物理功能 (PF) 配置相同的 Network Direct 技术。
- 每一 PF 最多有 16 个 VF 支持 VF RDMA。对于四个端口的适配器，每一 PF 最多有 8 个 VF。
- 仅 Windows Server 2019 （主机和 VM OS ）支持 VF RDMA。
- Windows Hypervisor 的 Linux VM 不支持 VF RDMA。
- NPAR 模式下不支持 VF RDMA。
- 每一 VF 最多支持 128 对队列对 (QP)/ 连接。
- 支持 PF 及其 VF 之间以及同一 PF 的各 VF 中的 RDMA 流量。这种流量模式称作*环回流量*。

- 在部分以前的服务器平台上，可能无法枚举 VF 设备中的 NIC PCI 的功能之一 (PF)。此限制是由于需新增 PCI 基本地址寄存器 （BAR）以支持 VF RDMA，这意味着 OS/BIOS 无法为每一 VF 分配所需的 BAR。
- 为了在一个 VM 中支持最多数量的 QP，假设仅向 VM 分配一个 VF，需约 8GB 的 RAM。如果向 VM 分配的 RAM 少于 8GB，由于内存不足和内存配置失败，可能会导致活动连接数量急剧减少。

在 Linux 的适配器上配置 RoCE

本节介绍 RHEL 和 SLES 的 RoCE 配置步骤。本节还介绍如何验证 RoCE 配置，并提供有关在 vLAN 接口使用组 ID (GID) 的一些指南。

RHEL 的 RoCE 配置

要在适配器上配置 RoCE，必须在 RHEL 主机上安装并配置开放结构企业分布 (OFED)。

为 RHEL 准备内建 OFED：

1. 在安装或升级操作系统时，选择 Infiniband 和 OFED 支持软件包。
2. 从 RHEL ISO 映像安装以下 RPM：

   `libibverbs-devel-x.x.x.x86_64.rpm`
   （libqedr 库所需）

   `perftest-x.x.x.x86_64.rpm`
   （InfiniBand 带宽和延迟应用程序所需）

   或使用 Yum，安装内建 OFED：

   ```
   yum groupinstall "Infiniband Support"
   yum install perftest
   yum install tcl tcl-devel tk zlib-devel libibverbs
   libibverbs-devel
   ```

注

在安装过程中，如果您已经选择了前面提到的软件包，则不需要重新安装它们。内建 OFED 和支持软件包可能因操作系统版本而异。

3. 如第 14 页上 “ 安装包含 RDMA 的 Linux 驱动程序 ”中所述，安装新 Linux 驱动程序。

SLES 的 RoCE 配置

要在 SLES 主机的适配器上配置 RoCE，必须在 SLES 主机上安装并配置 OFED。

要为 SLES 安装内建 OFED：

1. 在安装或升级操作系统时，选择 InfiniBand 支持软件包。
2. (SLES 12.*x*) 从相应的 SLES SDK 套件映像安装以下 RPM：

   `libibverbs-devel-x.x.x.x86_64.rpm`
   （libqedr 安装所需）

   `perftest-x.x.x.x86_64.rpm`
   （带宽和延迟应用程序所需）

3. (SLES 15/15 SP1) 安装以下 RPM。

   安装后，`rdma-core*`, `libibverbs*`, `libibumad*`, `libibmad*`, `librdmacm*`，和 `perftest` RPM 可能丢失 （都为 RDMA 安装所需软件包）。使用以下方法之一安装这些软件包：

   - ❑ 加载 Package DVD （DVD 软件包）并安装丢失的 RPM。
   - ❑ 使用 `zypper` 命令安装丢失的 RPM。例如：

     ```
     #zypper install rdma*
     #zypper install libib*
     #zypper install librdma*
     #zypper install perftest
     ```

4. 如第 14 页上 “ 安装包含 RDMA 的 Linux 驱动程序 ”中所述，安装 Linux 驱动程序。

验证 Linux 上的 RoCE 配置

安装 OFED 后，安装 Linux 驱动程序，然后加载 RoCE 驱动程序，验证在所有 Linux 操作系统上是否检测到 RoCE 设备。

要在 Linux 上验证 RoCE 配置：

1. 使用 `service/systemctl` 命令停止防火墙表。
2. 仅对于 RHEL：如果已安装 RDMA 服务 (`yum install rdma`)，请验证 RDMA 服务已启动。

注

对于 RHEL 7.*x* 和 SLES 12 SP*x* 及更高版本， RDMA 服务会在重新引导后自行启动。

在 RHEL 或 CentOS 上：使用 `service rdma` 状态命令启动服务：

- ❑ 如果 RDMA 尚未启动，请发出以下命令：

```
# service rdma start
```

- ❑ 如果 RDMA 尚未启动，则发出以下备选命令之一：

```
# /etc/init.d/rdma start
```

或者

```
# systemctl start rdma.service
```

3. 通过检查 dmesg 日志验证是否检测到 RoCE 设备：

```
# dmesg|grep qedr
  [87910.988411] qedr: discovered and registered 2 RoCE funcs
```

4. 验证是否已加载所有模块。例如：

```
# lsmod|grep qedr
  qedr                   89871  0
  qede                   96670  1 qedr
  qed                    2075255  2 qede,qedr
  ib_core                88311  16  qedr, rdma_cm, ib_cm,
                         ib_sa,iw_cm,xprtrdma,ib_mad,ib_srp,
                         ib_ucm,ib_iser,ib_srpt,ib_umad,
                         ib_uverbs,rdma_ucm,ib_ipoib,ib_isert
```

5. 使用配置方法 （如 ifconfig）配置 IP 地址并启用端口。例如：

```
# ifconfig ethX 192.168.10.10/24 up
```

6. 发出 `ibv_devinfo` 命令。对于每个 PCI 功能，您应该会看到一个单独的 `hca_id`，如下例中所示：

```
root@captain:~# ibv_devinfo
hca_id: qedr0
        transport:                      InfiniBand (0)
        fw_ver:                         8.3.9.0
        node_guid:                      020e:1eff:fe50:c7c0
        sys_image_guid:                 020e:1eff:fe50:c7c0
        vendor_id:                      0x1077
        vendor_part_id:                 5684
        hw_ver:                         0x0
        phys_port_cnt:                  1
                port:   1
                        state:                  PORT_ACTIVE (1)
                        max_mtu:                4096 (5)
                        active_mtu:             1024 (3)
                        sm_lid:                 0
                        port_lid:               0
                        port_lmc:               0x00
                        link_layer:             Ethernet
```

7. 验证所有服务器之间的 L2 和 RoCE 连接：一个服务器充当服务器，另一个服务器充当客户端。
   - ❑ 使用简单的 `ping` 命令验证 L2 连接。
   - ❑ 通过执行服务器或客户端的 RDMA ping 验证 RoCE 连接：

     在服务器上发出以下命令：

     `ibv_rc_pingpong -d <ib-dev> -g 0`

     在客户端上发出以下命令：

     `ibv_rc_pingpong -d <ib-dev> -g 0 <server L2 IP address>`

下面是服务器和客户端上的成功 ping pong 测试的示例。

服务器 Ping：

```
root@captain:~# ibv_rc_pingpong -d qedr0 -g 0
local address:  LID 0x0000, QPN 0xff0000, PSN 0xb3e07e, GID
fe80::20e:1eff:fe50:c7c0
remote address: LID 0x0000, QPN 0xff0000, PSN 0x934d28, GID
fe80::20e:1eff:fe50:c570
8192000 bytes in 0.05 seconds = 1436.97 Mbit/sec
1000 iters in 0.05 seconds = 45.61 usec/iter
```

客户端 Ping：

```
root@lambodar:~# ibv_rc_pingpong -d qedr0 -g 0 192.168.10.165
local address:  LID 0x0000, QPN 0xff0000, PSN 0x934d28, GID
fe80::20e:1eff:fe50:c570
remote address: LID 0x0000, QPN 0xff0000, PSN 0xb3e07e, GID
fe80::20e:1eff:fe50:c7c0
8192000 bytes in 0.02 seconds = 4211.28 Mbit/sec
1000 iters in 0.02 seconds = 15.56 usec/iter
```

- 要显示 RoCE 统计信息，请发出以下命令，其中 ***x*** 是设备号：

  > **mount -t debugfs nodev /sys/kernel/debug**
  > **cat /sys/kernel/debug/qedr/qedr*X*/stats**

vLAN 接口和 GID 索引值

如果您在服务器和客户端上均使用 vLAN 接口，还必须在交换机上配置相同的 vLAN ID。如果您是通过交换机运行流量，则 InfiniBand 应用程序必须使用正确的 GID 值，该值基于 vLAN ID 和 vLAN IP 地址。

根据以下结果，应该为任意 perftest 应用程序使用 GID 值 (-x 4/-x 5)。

```
# ibv_devinfo -d qedr0 -v|grep GID
     GID[  0]:  fe80:0000:0000:0000:020e:1eff:fe50:c5b0
     GID[  1]:  0000:0000:0000:0000:0000:ffff:c0a8:0103
     GID[  2]:  2001:0db1:0000:0000:020e:1eff:fe50:c5b0
     GID[  3]:  2001:0db2:0000:0000:020e:1eff:fe50:c5b0
     GID[  4]:  0000:0000:0000:0000:0000:ffff:c0a8:0b03
     GID[  5]:  fe80:0000:0000:0000:020e:1e00:0350:c5b0
```

vLAN 接口的 IP 地址

vLAN ID 3

注

背靠背或暂停设置的默认 GID 值为零 (0)。对于服务器和交换机配置，必须确定合适的 GID 值。如果使用的是交换机，请参阅相应的交换机配置说明文件以了解正确设置。

Linux 的 RoCE v2 配置

要验证 RoCE v2 功能，必须使用 RoCE v2 支持的内核。

为 Linux 配置 RoCE v2：

1. 请确保您使用以下支持的内核之一：
   - ❑ SLES 15/15 SP1

- ❑ SLES 12 SP4 及更高版本
- ❑ RHEL 7.6, 7.7 和 8.0

2. 执行以下操作配置 RoCE v2：

   a. 识别 RoCE v2 的 GID 索引。

   b. 配置服务器和客户端的路由地址。

   c. 在交换机上启用 L3 路由。

注

可通过使用 RoCE v2 支持的内核配置 RoCE v1 和 RoCE v2。可使用这些内核在同一子网以及不同子网 （例如 RoCE v2）和任何可路由环境中运行 RoCE 流量。 RoCE v2 只需少量设置，所有其他交换机和适配器设置均通用于 RoCE v1 和 v2。

识别 RoCE v2 GID 索引或地址

要查找 RoCE v1 和 v2 特定 GID，请使用 sys 或 class 参数，或从 41*xxx* FastLinQ 源文件包运行 RoCE 脚本。要检查默认 **RoCE GID 索引**和地址，请发出 `ibv_devinfo` 命令并将其与 sys 或 class 参数进行比较。例如：

```
#ibv_devinfo -d qedr0 -v|grep GID
            GID[  0]:           fe80:0000:0000:0000:020e:1eff:fec4:1b20
            GID[  1]:           fe80:0000:0000:0000:020e:1eff:fec4:1b20
            GID[  2]:           0000:0000:0000:0000:0000:ffff:1e01:010a
            GID[  3]:           0000:0000:0000:0000:0000:ffff:1e01:010a
            GID[  4]:           3ffe:ffff:0000:0f21:0000:0000:0000:0004
            GID[  5]:           3ffe:ffff:0000:0f21:0000:0000:0000:0004
            GID[  6]:           0000:0000:0000:0000:0000:ffff:c0a8:6403
            GID[  7]:           0000:0000:0000:0000:0000:ffff:c0a8:6403
```

使用 sys 和 class 参数验证 RoCE v1 或 RoCE v2 GID 索引和地址

使用以下任一选项，通过 sys 和 class 参数验证 RoCE v1 或 v2 GID 索引和地址：

- ■ **Option 1:**

```
# cat /sys/class/infiniband/qedr0/ports/1/gid_attrs/types/0
IB/RoCE v1
# cat /sys/class/infiniband/qedr0/ports/1/gid_attrs/types/1
RoCE v2

# cat /sys/class/infiniband/qedr0/ports/1/gids/0
fe80:0000:0000:0000:020e:1eff:fec4:1b20
```

```
# cat /sys/class/infiniband/qedr0/ports/1/gids/1
fe80:0000:0000:0000:020e:1eff:fec4:1b20
```

- **选项 2：**

  使用 FastLinQ 源文件包的脚本。

```
#/../fastlinq-8.x.x.x/add-ons/roce/show_gids.sh
DEV    PORT  INDEX  GID                                     IPv4           VER     DEV
---    ----  -----  ---                                     ------------   ---     ---
qedr0  1     0      fe80:0000:0000:0000:020e:1eff:fec4:1b20                v1      p4p1
qedr0  1     1      fe80:0000:0000:0000:020e:1eff:fec4:1b20                v2      p4p1
qedr0  1     2      0000:0000:0000:0000:0000:ffff:1e01:010a  30.1.1.10     v1      p4p1
qedr0  1     3      0000:0000:0000:0000:0000:ffff:1e01:010a  30.1.1.10     v2      p4p1
qedr0  1     4      3ffe:ffff:0000:0f21:0000:0000:0000:0004                v1      p4p1
qedr0  1     5      3ffe:ffff:0000:0f21:0000:0000:0000:0004                v2      p4p1
qedr0  1     6      0000:0000:0000:0000:0000:ffff:c0a8:6403  192.168.100.3  v1      p4p1.100
qedr0  1     7      0000:0000:0000:0000:0000:ffff:c0a8:6403  192.168.100.3  v2      p4p1.100
qedr1  1     0      fe80:0000:0000:0000:020e:1eff:fec4:1b21                v1      p4p2
qedr1  1     1      fe80:0000:0000:0000:020e:1eff:fec4:1b21                v2      p4p2
```

注

您必须基于服务器或交换机配置 (暂停 /PFC) 指定 RoCE v1 或 v2 的 GID 索引值。使用链路本地 Ipv6 地址、 IPv4 地址或 Ipv6 地址的 GID 索引。要将带 vLAN 标记的帧用于 RoCE 流量，您必须指定从 vLAN IPv4 或 IPv6 地址得出的 GID 索引值。

通过 perftest 应用程序验证 RoCE v1 或 v2 功能

本节展示如何通过 perftest 应用程序验证 RoCE v1 或 v2 功能。在本示例中，使用以下服务器 IP 和客户端 IP：

- 服务器 IP:192.168.100.3
- 客户端 IP:192.168.100.4

验证 RoCE v1

在同一子网上运行并使用 RoCE v1 GID 索引。

```
Server# ib_send_bw -d qedr0 -F -x 0
Client# ib_send_bw -d qedr0 -F -x 0  192.168.100.3
```

验证 RoCE v2

在同一子网上运行并使用 RoCE v2 GID 索引。

```
Server# ib_send_bw -d qedr0 -F -x 1
Client# ib_send_bw -d qedr0 -F -x 1  192.168.100.3
```

注

如果通过交换机 PFC 配置运行，请对通过同一子网的 RoCE v1 或 v2 使用 vLAN GID。

通过不同子网验证 RoCE v2

注

您必须首先配置交换机和服务器的路由设置。通过 HII、 UEFI 用户界面或其中一个 Marvell 管理公用程序在适配器上设置 RoCE 优先级和 DCBX 模式。

通过不同子网验证 RoCE v2：

1. 使用 DCBX-PFC 配置设置服务器和客户端的路由配置。

   - **系统设置：**

     服务器 VLAN IP：192.168.100.3 和**网关**：192.168.100.1

     客户端 VLAN IP：192.168.101.3 和**网关**：192.168.101.1

   - **服务器配置：**

```
#/sbin/ip link add link p4p1 name p4p1.100 type vlan id 100
#ifconfig p4p1.100 192.168.100.3/24 up
#ip route add 192.168.101.0/24 via 192.168.100.1 dev p4p1.100
```

   - **客户端配置：**

```
#/sbin/ip link add link p4p1 name p4p1.101 type vlan id 101
#ifconfig p4p1.101 192.168.101.3/24 up
#ip route add 192.168.100.0/24 via 192.168.101.1 dev p4p1.101
```

2. 使用以下程序设置交换机设置。

   - 使用任何流控制方法 （暂停、 DCBX-CEE 或 DCBX-IEEE），并为 RoCE v2 启用 IP 路由。请参阅第 128 页上 “ 准备以太网交换机 ” 以了解 RoCE v2 配置，或参阅供应商交换机文档。

- 如果您使用 PFC 配置和 L3 路由，则在使用不同子网的 vLAN 上运行 RoCE v2 流量，并使用 RoCE v2 vLAN GID 索引。

```
Server# ib_send_bw -d qedr0 -F -x 5
Client# ib_send_bw -d qedr0 -F -x 5 192.168.100.3
```

服务器交换机设置：

```
[root@RoCE-Auto-2 /]# ib_send_bw -d qedr0 -F -x 5 -q 2 --report_gbits

************************************
* Waiting for client to connect... *
************************************
---------------------------------------------------------------------------------------
                    Send BW Test
 Dual-port       : OFF          Device         : qedr0
 Number of qps   : 2            Transport type : IB
 Connection type : RC           Using SRQ      : OFF
 RX depth        : 512
 CQ Moderation   : 100
 Mtu             : 1024[B]
 Link type       : Ethernet
 Gid index       : 5
 Max inline data : 0[B]
 rdma_cm QPs     : OFF
 Data ex. method : Ethernet
---------------------------------------------------------------------------------------
 local address: LID 0000 QPN 0xff0000 PSN 0xf0b2c3
 GID: 00:00:00:00:00:00:00:00:00:00:255:255:192:168:100:03
 local address: LID 0000 QPN 0xff0002 PSN 0xa2b8f1
 GID: 00:00:00:00:00:00:00:00:00:00:255:255:192:168:100:03
 remote address: LID 0000 QPN 0xff0000 PSN 0x40473a
 GID: 00:00:00:00:00:00:00:00:00:00:255:255:192:168:101:03
 remote address: LID 0000 QPN 0xff0002 PSN 0x124cd3
 GID: 00:00:00:00:00:00:00:00:00:00:255:255:192:168:101:03
---------------------------------------------------------------------------------------
 #bytes     #iterations    BW peak[Gb/sec]    BW average[Gb/sec]   MsgRate[Mpps]
 65536      1000             0.00               23.07              0.043995
---------------------------------------------------------------------------------------
```

图 7-12. 交换机设置，服务器

客户端交换机设置：

```
[root@roce-auto-1 ~]# ib_send_bw -d qedr0 -F -x 5 192.168.100.3 -q 2 --report_gbits
---------------------------------------------------------------------------------------
                    Send BW Test
 Dual-port       : OFF          Device         : qedr0
 Number of qps   : 2            Transport type : IB
 Connection type : RC           Using SRQ      : OFF
 TX depth        : 128
 CQ Moderation   : 100
 Mtu             : 1024[B]
 Link type       : Ethernet
 Gid index       : 5
 Max inline data : 0[B]
 rdma_cm QPs     : OFF
 Data ex. method : Ethernet
---------------------------------------------------------------------------------------
 local address: LID 0000 QPN 0xff0000 PSN 0x40473a
 GID: 00:00:00:00:00:00:00:00:00:00:255:255:192:168:101:03
 local address: LID 0000 QPN 0xff0002 PSN 0x124cd3
 GID: 00:00:00:00:00:00:00:00:00:00:255:255:192:168:101:03
 remote address: LID 0000 QPN 0xff0000 PSN 0xf0b2c3
 GID: 00:00:00:00:00:00:00:00:00:00:255:255:192:168:100:03
 remote address: LID 0000 QPN 0xff0002 PSN 0xa2b8f1
 GID: 00:00:00:00:00:00:00:00:00:00:255:255:192:168:100:03
---------------------------------------------------------------------------------------
 #bytes     #iterations    BW peak[Gb/sec]    BW average[Gb/sec]   MsgRate[Mpps]
 65536      1000             23.04              23.04              0.043936
---------------------------------------------------------------------------------------
```

图 7-13. 交换机设置，客户端

为 RDMA_CM 应用程序配置 RoCE v1 或 v2 设置

使用以下 FastLinQ 源文件包的脚本配置 RoCE:

```
# ./show_rdma_cm_roce_ver.sh
qedr0 is configured to IB/RoCE v1
qedr1 is configured to IB/RoCE v1

# ./config_rdma_cm_roce_ver.sh v2
configured rdma_cm for qedr0 to RoCE v2
configured rdma_cm for qedr1 to RoCE v2
```

服务器设置：

```
[root@RoCE-Auto-2 /]# rping -s -v -C 10
server ping data: rdma-ping-0: ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqr
server ping data: rdma-ping-1: BCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrs
server ping data: rdma-ping-2: CDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrst
server ping data: rdma-ping-3: DEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstu
server ping data: rdma-ping-4: EFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuv
server ping data: rdma-ping-5: FGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvw
server ping data: rdma-ping-6: GHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwx
server ping data: rdma-ping-7: HIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxy
server ping data: rdma-ping-8: IJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz
server ping data: rdma-ping-9: JKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyzA
server DISCONNECT EVENT...
wait for RDMA_READ_ADV state 10
[root@RoCE-Auto-2 /]#
```

图 7-14. 配置 RDMA_CM 应用程序：服务器

客户端设置：

```
[root@roce-auto-1 ~]# rping -c -v -C 10 -a 192.168.100.3
ping data: rdma-ping-0: ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqr
ping data: rdma-ping-1: BCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrs
ping data: rdma-ping-2: CDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrst
ping data: rdma-ping-3: DEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstu
ping data: rdma-ping-4: EFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuv
ping data: rdma-ping-5: FGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvw
ping data: rdma-ping-6: GHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwx
ping data: rdma-ping-7: HIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxy
ping data: rdma-ping-8: IJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz
ping data: rdma-ping-9: JKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyzA
client DISCONNECT EVENT...
[root@roce-auto-1 ~]#
```

图 7-15. 配置 RDMA_CM 应用程序：客户端

为 SR-IOV VF 设备 (VF RDMA) 配置 RoCE

以下章节介绍如何在 Linux 上为 SR-IOV VF 设备 （也称作 VF RDMA）配置 RoCE。同时也提供相关信息和限制。

表 7-4 列出了支持的 Linux 操作系统组合。

表 7-4. 支持 VF RDMA 的 Linux 操作系统

	来宾账户 OS					
虚拟机监控程序	RHEL 7.6	RHEL 7.7	RHEL 8.0	SLES12 SP4	SLES15 SP0	SLES15 SP1
	是	是	是	是	是	是
RHEL 7.7	是	是	是	是	是	是
RHEL 8.0	是	是	是	是	是	是
SLES12 SP4	是	是	是	是	是	是
SLES15 SP0	是	是	是	是	是	是
SLES15 SP1	是	是	是	是	是	是

如果您正在使用内建 OFED，请在虚拟机监控程序主机操作系统和来宾 (VM) 操作系统之间使用相同的 OFED 分布。查看开箱即用的 OFED 分布发行说明，了解其特定支持的主机操作系统到虚拟机操作系统的分布矩阵。

枚举 L2 和 RDMA 的 VF

枚举 VF 有两种方法：

- 用户定义的 VF MAC 分配
- 动态或随机的 VF MAC 分配

用户定义的 VF MAC 分配

定义 VF MAC 分配时，默认 VF 枚举方法没有变化。创建 VF 数量后，分配静态 MAC 地址。

要创建用户定义的 VF MAC 分配：

1. 枚举默认 VF。

```
# modprobe -v qede
# echo 2 > /sys/class/net/p6p1/device/sriov_numvfs
# ip link show
14: p6p1: <NO-CARRIER,BROADCAST,MULTICAST,UP> mtu 1500 qdisc mq state DOWN
mode DEFAULT group default qlen 1000
    link/ether 14:02:ec:ce:d0:e4 brd ff:ff:ff:ff:ff:ff
    vf 0 MAC 00:00:00:00:00:00, spoof checking off, link-state auto
    vf 1 MAC 00:00:00:00:00:00, spoof checking off, link-state auto
```

2. 分配静态 MAC 地址：

```
#  ip link set dev p6p1 vf 0 mac 3c:33:44:55:66:77
#  ip link set dev p6p1 vf 1 mac 3c:33:44:55:66:89
#ip link show
14: p6p1: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 1500 qdisc mq state UP mode DEFAULT group default
qlen 1000
    link/ether 14:02:ec:ce:d0:e4 brd ff:ff:ff:ff:ff:ff
    vf 0 MAC 3c:33:44:55:66:77, tx rate 25000 (Mbps), max_tx_rate 25000Mbps, spoof checking off,
link-state auto
    vf 1 MAC 3c:33:44:55:66:89, tx rate 25000 (Mbps), max_tx_rate 25000Mbps, spoof checking off,
link-state auto
```

3. 为了反映 RDMA 的情况，如果 `qedr` 驱动程序已经加载，请重新加载。

```
#rmmod qedr
#modprobe        qedr
#ibv_devices
    device                 node GUID
    ------              ----------------
    qedr0               1602ecfffeced0e4
    qedr1               1602ecfffeced0e5
```

```
    qedr_vf0            3e3344fffe556677
    qedr_vf1            3e3344fffe556689
```

动态或随机的 VF MAC 分配

要动态分配 VF MAC：

```
# modprobe -r qedr
# modprobe -v qed vf_mac_origin=3 [使用此模块参数进行动态 MAC 分配]
# modprobe -v qede
# echo 2 > /sys/class/net/p6p1/device/sriov_numvfs
# modprobe qedr (This is an optional, mostly qedr driver loads
itself)
# ip link show|grep vf
    vf 0 MAC ba:1a:ad:08:89:00, tx rate 25000 (Mbps), max_tx_rate
25000Mbps, spoof checking off, link-state auto
    vf 1 MAC 96:40:61:49:cd:68, tx rate 25000 (Mbps), max_tx_rate
25000Mbps, spoof checking off, link-state auto
# lsmod |grep qedr
# ibv_devices
    device                 node GUID
    ------              ----------------
    qedr0               1602ecfffececfa0
    qedr1               1602ecfffececfa1
    qedr_vf0            b81aadfffe088900
    qedr_vf1            944061fffe49cd68
```

支持 RDMA 的 VF 数量

对于 41*xxx* 系列适配器， L2 和 RDMA 的 VF 数量是根据可用资源共享的。

双端口适配器

每个 PF 最多支持 RDMA 40 个 VF ；如果 VF 的数量超过 56，它将被 VF 的总数减去 (96)。

在以下示例中， PF0 为 `/sys/class/net/<PF-interface>/device/sriov_numvfs`

`Echo 40 > PF0` （L2+RDMA 的 VF=40+40 （L2 和 RDMA 均可使用 40 个 VF））
`Echo 56 > PF0` （L2+RDMA 的 VF=56+40）

超出 56 个 VF 后，该数值将被 VF 总数减去。例如：

`echo 57 > PF0 then 96-57=39 VFs for RDMA` （L2 的 57 个 VF + RDMA 的 39 个 VF）
`echo 96 > PF0 then 96-96=0 VFs for RDMA` （所有 96 个 VF 只能用于 L2）

要查看 L2 和 RDMA 的可用 VF：

```
L2        : # ip link show
RDMA: # ibv_devices
```

四端口适配器

每个 PF 最多支持 RDMA 20 个 VF ；直到 48 个 Vf，此时 RDMA 有 20 个 VF。当超过 28 个 VF 时，该数值将被总 VF 数减去 (48)。

例如，在 4x10G 中：

`Echo 20 > PF0` （L2+RDMA 的 VF=20+20）
`Echo 28 > PF0` （L2+RDMA 的 VF=28+20）

当超过 28 个 VF 时，该数值将被 VF 总数减去。例如：

`echo 29 > PF0` （48-29=RDMA 的 19 个 VF ； L2 的 29 个 VF + RDMA 的 19 个 VF）
`echo 48 > PF0` （48-48=RDMA 的 0 个 VF ；所有 48 个 VF 只能用于 L2）

限制

VF RDMA 具有以下限制：

- 不支持 iWARP
- 不支持 NPAR
- 不支持跨操作系统；例如， Linux 虚拟机监控程序不能使用 Windows 来宾操作系统 (VM)
- VF 接口上的 Perftest 延迟测试只能使用内嵌大小 0 `-I 0` 选项运行。默认值和多个内嵌大小都无法运行。
- 要允许 RDMA_C M 应用程序在不同于默认大小 (1500) 的 MTU 大小 (512–9000) 上运行，请执行以下步骤：

  1. 卸载 `qedr` 驱动程序：

     `#rmmod qedr`

  2. 在 VF 接口上设置 MTU：

     `#ifconfig <VF interface> mtu 9000`

  3. 加载 `qedr` 驱动程序：

     `#modprobe qedr`

- `Rdma_server/rdma_xserver` 不支持 VF 接口。
- VF 上不支持 RDMA 绑定

在 VMware ESX 的适配器上配置 RoCE

本节提供以下 RoCE 配置的步骤和信息：

- 配置 RDMA 接口
- 配置 MTU
- RoCE 模式和统计信息
- 配置半虚拟化 RDMA 设备 (PVRDMA)

注

向 RDMA 速度映射以太网速度并不总是准确，因为 RoCE 驱动程序可指定的值与 Infiniband® 对应。

例如，如果在操作速度为 1Gbps 的以太网接口上配置 RoCE，此时 RDMA 的速度显示为 2.5Gbps。在 ESXi 提供的头文件中没有其它合适的值可用于 RoCE 驱动程序显示 1Gbps 的速度。

配置 RDMA 接口

配置 RDMA 接口：

1. 安装 Marvell NIC 和 RoCE 驱动程序。
2. 使用模块参数，通过发出以下命令从 NIC 驱动程序启用 RoCE 功能：

   esxcfg-module -s 'enable_roce=1' qedentv

   如需应用更改，重新加载 NIC 驱动程序或重新引导系统。
3. 如需查看 NIC 接口列表，请发出 `esxcfg-nics -l` 命令。例如：

```
esxcfg-nics -l
Name    PCI           Driver      Link Speed      Duplex MAC Address       MTU    Description
Vmnic0  0000:01:00.2 qedentv     Up   25000Mbps  Full   a4:5d:36:2b:6c:92 1500   QLogic Corp.
QLogic FastLinQ QL41xxx 1/10/25 GbE Ethernet Adapter
Vmnic1  0000:01:00.3 qedentv     Up   25000Mbps  Full   a4:5d:36:2b:6c:93 1500   QLogic Corp.
QLogic FastLinQ QL41xxx 1/10/25 GbE Ethernet Adapter
```

4. 如需查看 RDMA 设备列表，请发出 `esxcli rdma device list` 命令。例如：

```
esxcli rdma device list
Name     Driver   State    MTU  Speed    Paired Uplink  Description
-------  -------  ------  ----  -------  -------------  ------------------------------
vmrdma0  qedrntv  Active  1024  25 Gbps  vmnic0     QLogic FastLinQ QL45xxx RDMA Interface
vmrdma1  qedrntv  Active  1024  25 Gbps  vmnic1     QLogic FastLinQ QL45xxx RDMA Interface
```

5. 如需创建新虚拟交换机，请发出以下命令：

   esxcli network vswitch standard add -v <new vswitch name>

   例如：

   # **esxcli network vswitch standard add -v roce_vs**

   此操作创建了一个名为 *roce_vs* 的新虚拟交换机。

6. 如需连接 Marvell NIC 端口至 vSwitch，请发出以下命令：

   # **esxcli network vswitch standard uplink add -u <uplink device> -v <roce vswitch>**

   例如：

   # **esxcli network vswitch standard uplink add -u vmnic0 -v roce_vs**

7. 如需在该 vSwitch 上创建新端口组，请发出以下命令：

   # **esxcli network vswitch standard portgroup add -p roce_pg -v roce_vs**

   例如：

   # **esxcli network vswitch standard portgroup add -p roce_pg -v roce_vs**

8. 如需在该端口组创建一个 vmknic 接口并配置 IP，请发出以下命令：

   # **esxcfg-vmknic -a -i <IP address> -n <subnet mask> <roce port group name>**

   例如：

   # **esxcfg-vmknic -a -i 192.168.10.20 -n 255.255.255.0 roce_pg**

9. 如需配置 vLAN ID，请发出以下命令：

   # **esxcfg-vswitch -v <VLAN ID> -p roce_pg**

   如需通过 vLAN ID 运行 RoCE 流量，请配置对应 VMkernel 端口组上的 vLAN ID。

配置 MTU

如需修改 RoCE 接口的 MTU，请更改相应 vSwitch 的 MTU。基于 vSwitch 的 MTU，通过发出以下命令设置 RDMA 接口的 MTU 大小：

esxcfg-vswitch -m <new MTU> <RoCE vswitch name>

例如：

```
# esxcfg-vswitch -m 4000 roce_vs
# esxcli rdma device list
Name     Driver   State    MTU  Speed    Paired Uplink  Description
-------  -------  ------  ----  -------  -------------  ------------------------------
vmrdma0  qedrntv  Active     2048   25 Gbps  vmnic0    QLogic FastLinQ QL45xxx RDMA Interface
vmrdma1  qedrntv  Active     1024   25 Gbps  vmnic1    QLogic FastLinQ QL45xxx RDMA Interface
```

RoCE 模式和统计信息

对于 RoCE 模式，ESXi 要求 RoCE v1 和 v2 的共存支持。队列对创建期间确定需使用的 RoCE 模式。在注册和初始化期间可在 ESXi 驱动器中选择两种模式。要查看 RoCE 统计信息，请发出以下命令：

```
# esxcli rdma device stats get -d vmrdma0
   Packets received: 0
   Packets sent: 0
   Bytes received: 0
   Bytes sent: 0
   Error packets received: 0
   Error packets sent: 0
   Error length packets received: 0
   Unicast packets received: 0
   Multicast packets received: 0
   Unicast bytes received: 0
   Multicast bytes received: 0
   Unicast packets sent: 0
   Multicast packets sent: 0
   Unicast bytes sent: 0
   Multicast bytes sent: 0
   Queue pairs allocated: 0
   Queue pairs in RESET state: 0
   Queue pairs in INIT state: 0
   Queue pairs in RTR state: 0
   Queue pairs in RTS state: 0
   Queue pairs in SQD state: 0
   Queue pairs in SQE state: 0
   Queue pairs in ERR state: 0
   Queue pair events: 0
   Completion queues allocated: 1
   Completion queue events: 0
```

```
Shared receive queues allocated: 0
Shared receive queue events: 0
Protection domains allocated: 1
Memory regions allocated: 3
Address handles allocated: 0
Memory windows allocated: 0
```

配置半虚拟化 RDMA 设备 (PVRDMA)

请参阅 VMware 文档 （例如，https://kb.vmware.com/articleview?docid=2147694）查看通过 vCenter 接口配置 PVRDMA 的详细信息。以下说明仅供参考。

使用 vCenter 接口配置 PVRDMA：

1. 按照以下方式创建和配置新分布式虚拟交换机：
   a. 在 VMware vSphere® Web 客户端，右键单击 Navigator （导航）窗口左侧窗格中的 **RoCE** 节点。
   b. 在 Actions （活动）菜单上，指向 **Distributed Switch** （分布式交换机），然后单击 **New Distributed Switch** （新分布式交换机）。
   c. 选择 6.5.0 版本。
   d. 在 **New Distributed Switch** （新分布式交换机）下，单击 **Edit settings** （编辑设置），并配置以下：
      - **Number of uplinks**。选择正确值。
      - **Network I/O Control**。选择 **Disabled** （已禁用）。
      - **Default port group**。选择 **Create a default port group** （创建默认端口组）复选框。
      - **Port group name**。键入端口组名称。

      图 7-16 显示一个示例。

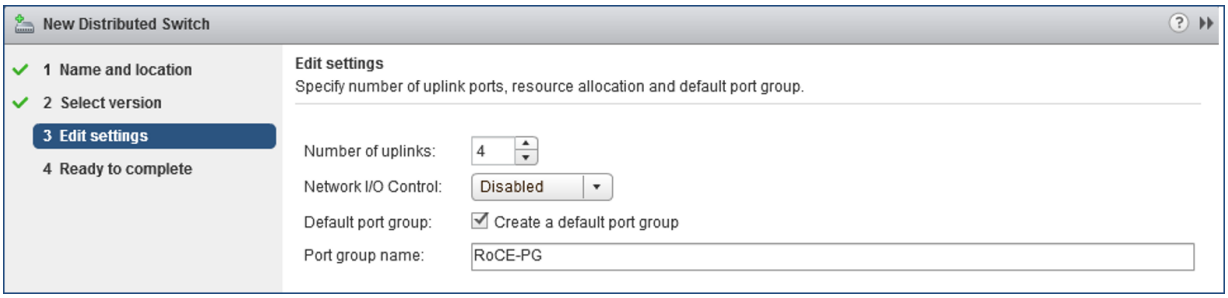


图 7-16. 配置新分布式交换机

2. 执行以下步骤配置分布式虚拟交换机：
   a. 在 VMware vSphere Web 客户端，扩展 Navigator （导航）窗口左侧窗格中的 **RoCE** 节点。
   b. 右键单击 **RoCE-VDS**，然后单击 **Add and Manage Hosts** （添加和管理主机）。
   c. 在 **Add and Manage Hosts** （添加和管理主机）下，进行以下配置：
      - **Assign uplinks**。从可用上行链路列表中选择。
      - **Manage VMkernel network adapters**。接受默认，然后单击 **Next** （下一步）。
      - **Migrate VM networking**。分配在 步骤 1 中创建的端口组。
3. 分配 PVRDMA 的 vmknic 用于 ESX 主机：
   a. 右键单击主机，然后单击 **Settings** （设置）。
   b. 在 Settings （设置）页面上，扩展 **System** （系统）字节，然后单击 **Advanced System Settings** （高级系统设置）。
   c. Advanced System Settings （高级系统设置）页面显示密匙对值及其摘要。单击 **Edit** （编辑）。
   d. 在 Edit Advanced System Settings （编辑高级系统设置）页面上，通过 **PVRDMA** 过滤以缩减所有设置刚好至 Net.PVRDMAVmknic。
   e. 设置 **Net.PVRDMAVmknic** 值至 **vmknic** ；例如， **vmk1**。

      图 7-17 显示一个示例。

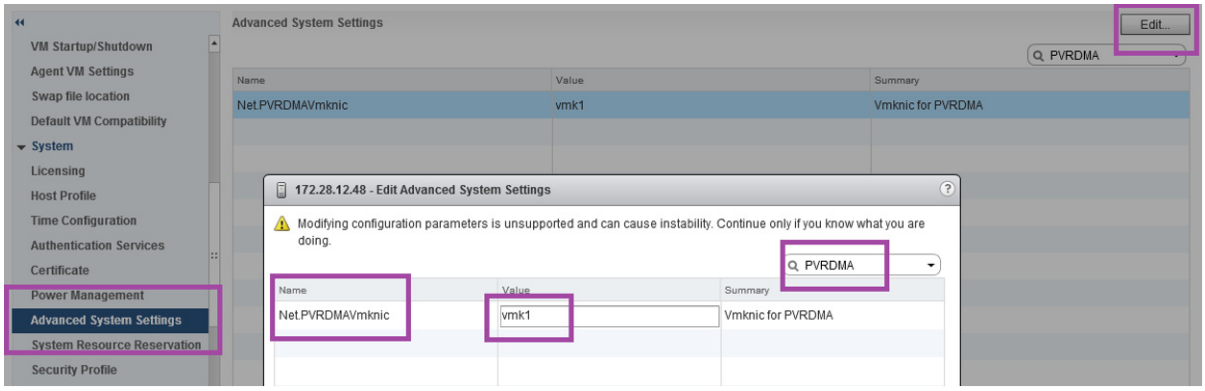


图 7-17. 为 PVRDMA 分配 vmknic

4. 设置 PVRDMA 防火墙规则：

   a. 右键单击主机，然后单击 **Settings** （设置）。

   b. 在 Settings （设置）页面上，扩展 **System** （系统）字节，然后单击 **Security Profile** （安全配置文件）。

   c. 在 Firewall Summary （防火墙摘要）页面上，单击 **Edit** （编辑）。

   d. 在 Edit Security Profile （编辑安全配置文件）对话框的 **Name** （名称）下，下拉选择 **pvrdma** 复选框，然后选择 **Set Firewall** （设置防火墙）复选框。

   图 7-18 显示一个示例。

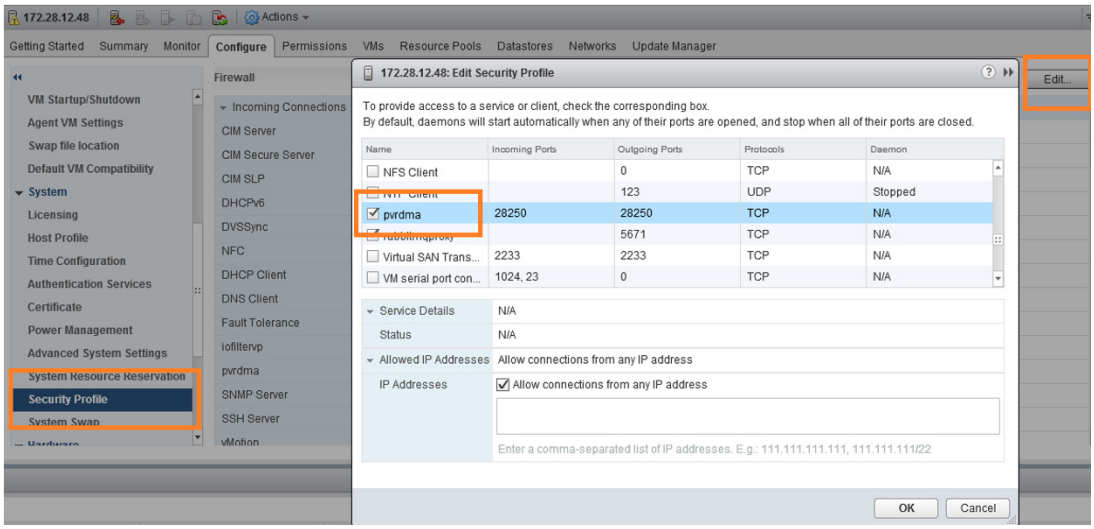


图 7-18. 设置防火墙规则

5. 执行以下步骤设置 PVRDMA 的 VM：

   a. 安装以下支持的来宾账户 OS：

      - RHEL 7.5, 7.6 和 8.0

   b. 安装 OFED4.17-1。

   c. 编译并安装 PVRDMA 来宾账户驱动程序和库。

   d. 执行以下步骤添加一个新的 PVRDMA 网络适配器至 VM：

      - 编辑 VM 设置。
      - 添加新的网络适配器。
      - 选择新添加的 DVS 端口组为 **Network** （网络）。
      - 选择 **PVRDMA** 为适配器类型。

   e. 引导 VM 后，请确保 PVRDMA 来宾账户驱动程序已加载。

配置 DCQCN

数据中心量化拥塞通知 (DCQCN) 为确定 RoCE 接收器如何通知传输器它们之间的交换机已提供一则显式拥塞通知 （通知点），以及传输器如何对此类通知作出反应（反应点）的功能。

本节提供以下有关 DCQCN 的配置信息：

DCQCN 术语

以下术语描述 DCQCN 配置：

- **ToS** （服务类型）为 IPv4 标头字段中的单字节。 ToS 包含：两个 ECN 最低有效位 (LSB) 和六个差分服务代码点 (DSCP) 最高有效位 (MSB)。 IPv6 流量级等同于 IPv4 ToS。
- **ECN** （显式拥塞通知）是交换机添加一则即将拥塞的指示到对外输出流量的机制。
- **CNP** （拥塞通知数据包）是通知点用于指示来自交换机的 ECN 回到反应点的数据包。 *InfiniBand Architecture Specification Volume 1 Release 1.2.1（InfiniBand 体系结构说明第 1.2.1 版第 1 卷）*补充中定义了 CNP，网址为：

  https://cw.infinibandta.org/document/dl/7781
- **VLAN Priority** （VLAN 优先级）是 L2 vLAN 标头中的一节字段。该字段在 vLAN 标签中为三个 MSB。
- **PFC** （基于优先级的流控制）为适用于承载特定 vLAN 优先级的流量的一种流控制机制。
- **DSCP-PFC** 为允许接收器出于 PFC 目的解释即将到来的数据包的优先级的功能，而非按照 vLAN 优先级或 IPv4 标头中的 DSCP 字段。您可使用间接表指示特定的 DSCP 值作为 vLAN 优先级值。因为 DSCP-PFC 为一种 L3 (IPv4) 功能，故它可在 L2 网络中运行。

- **Traffic classes** （流量类），也称作优先级组，是可具有有损或无损属性的 vLAN 优先级组 （或者如使用 DSCP-PFC 时，则为 DSCP 值）。通常， 0 用于默认有损流量类，3 用于 FCoE 流量类，4 用于 iSCSI-TLV 流量类。如果您尝试在同时支持 **FCoE** 或 iSCSI-TLV 流量的网络再次使用这些数字，您可能会遇到 DCB 错误匹配问题。Marvell 建议您对 RoCE 相关的流量类使用 1–2 或 5–7 中的数字。
- **ETS** （增强的转换服务）是每一流量级的最大网宽分配。

DCQCN 概览

有些网络协议 （例如 RoCE）要求无丢包。 PFC 是一种在 L2 网络中实现无丢包的机制，而 DSCP-PFC 是在不同 L2 网络中实现无丢包的机制。但 PFC 在以下方面存在不足：

- 当激活时， PFC 完全停止了端口上指定优先级的流量，而不是降低传输速率。
- 即使存在导致拥塞的特定连接子集，指定优先级的所有流量也会受影响。
- PFC 为一种单跃点机制。也就是说，如果接收者遇到拥塞并通过 PFC 数据包指出拥塞，只有最邻近的设备会反应。当邻近设备遇到拥塞 （可能是由于它无法再传输）时，也会生成自己的 PFC。这种生成也称作 *暂停传播*。暂停传播可能造成路由利用率低下，因为必须等到所有缓冲区都拥塞之后传输者才发现问题。

DCQCN 可解决此类所有缺陷。 ECN 向反应点提供拥塞指示。反应点向传输器发送 CNP 数据包，传输器通过降低其传输速率并避免拥塞作出反应。 DCQCN 也指定传输器如何在拥塞结束后尝试提高其传输速率和有效使用带宽。 2015 SIGCOMM 文件中的 *Congestion Control for Large-Scale RDMA Deployments*（*大规模 RDMA 部署的拥塞控制*）介绍了 DCQCN，网址为：

http://conferences.sigcomm.org/sigcomm/2015/pdf/papers/p523.pdf

DCB 相关的参数

使用 DCB 向流量级 （优先级组）映射优先级。 DCB 也控制受限于 PFC 的优先级组 （无损流量）和相关带宽分配 (ETS)。

RDMA 流量上的全局设置

RDMA 流量上的全局设置包括配置 vLAN 优先级、 ECN 和 DSCP。

设置 RDMA 流量的 vLAN 优先级

创建 QP （队列对）时使用应用程序设置指定 RDMA QP 使用的 vLAN 优先级。例如， `ib_write_bw` 基准控制优先级是使用 `-sl` 参数。当 RDMA-CM （RDMA 通信管理器）存在时，您可能无法设置优先级。

控制 vLAN 优先级的另一种方法是使用 rdma_glob_vlan_pri 节点。这种方法影响设置值后创建的 QP。例如，如需为随后创建的 QP 设置 vLAN 优先级数字为 5，请发出以下命令：

```
./debugfs.sh -n eth0 -t rdma_glob_vlan_pri 5
```

在 RDMA 流量上设置 ECN

使用 rdma_glob_ecn 节点启用指定 RoCE 优先级的 ECN。例如，如需使用优先级 5 启用 RoCE 流量的 ECN，请发出以下命令：

```
./debugfs.sh -n eth0 -t rdma_glob_ecn 1
```

通常在启用 DCQCN 后需要此命令。

在 RDMA 流量上设置 DSCP

使用 rdma_glob_dscp 节点控制 DSCP。例如，如需使用优先级 5 设置 RoCE 流量上的 DSCP，请发出以下命令：

```
./debugfs.sh -n eth0 -t rdma_glob_dscp 6
```

通常在启用 DCQCN 后需要此命令。

配置 DSCP-PFC

使用 dscp_pfc 节点配置 PFC 的 dscp->priority 关联。您必须在添加条目至映射前启用此功能。例如，如需映射 DSCP 值 6 至优先级 5，请发出以下命令：

```
./debugfs.sh -n eth0 -t dscp_pfc_enable 1
./debugfs.sh -n eth0 -t dscp_pfc_set 6 5
```

启用 DCQCN

如需启用 RoCE 流量的 DCQCN，请通过 dcqcn_enable 模块参数探测 qed 驱动程序。 DCQCN 要求启用 ECN 指示 （请参阅第 171 页上 “ 在 RDMA 流量上设置 ECN”）。

配置 CNP

可对拥塞通知数据包 (CNP) 独立配置 vLAN 优先级和 DSCP。使用 dcqcn_cnp_dscp 和 dcqcn_cnp_vlan_priority 模块参数控制此类数据包。例如：

```
modprobe qed dcqcn_cnp_dscp=10 dcqcn_cnp_vlan_priority=6
```

DCQCN 算法参数

表 7-5 列出 DCQCN 的算法参数。

表 7-5. DCQCN 算法参数

参数	描述和值
dcqcn_cnp_send_timeout	各 CNP 之间的发送时间的最小差异。单位为微秒。值的范围为 50..500000。
dcqcn_cnp_dscp	用于 CNP 的 DSCP 值。值的范围为 0..63。
dcqcn_cnp_vlan_priority	用于 CNP 的 vLAN 优先级。值的范围为 0..7。FCoE-Offload （FCoE 卸载）使用 3 而 iSCSI-Offload-TLV 通常使用 4。 Marvell 建议您指定 1-2 或 5-7 中的一个数值。在整个网络中使用这一相同值。
dcqcn_notification_point	0 – 禁用 DCQCN 通知点。 1 – 启用 DCQCN 通知点。
dcqcn_reaction_point	0 – 禁用 DCQCN 反应点。 1 – 启用 DCQCN 反应点。
dcqcn_rl_bc_rate	字节计数器限制
dcqcn_rl_max_rate	最大速率 (Mbps)
dcqcn_rl_r_ai	有效增加速率 (Mbps)
dcqcn_rl_r_hai	超有效增加速率 (Mbps)
dcqcn_gd	Alpha 更新增益要素。为 1/32 设置为 32，依此类推。
dcqcn_k_us	Alpha 更新间隔
dcqcn_timeout_us	DCQCN 超时

MAC 统计信息

需查看 MAC 统计信息，包括各优先级的 PFC 统计信息，请发出 phy_mac_stats 命令。例如，如需查看端口 1 的统计信息，请发出以下命令：

```
./debugfs.sh -n eth0 -d phy_mac_stat -P 1
```

脚本示例

以下示例可用作脚本：

```
# probe the driver with both reaction point and notification point enabled
# with cnp dscp set to 10 and cnp vlan priority set to 6
modprobe qed dcqcn_enable=1 dcqcn_notification_point=1 dcqcn_reaction_point=1
dcqcn_cnp_dscp=10 dcqcn_cnp_vlan_priority=6
modprobe qede

# dscp-pfc configuration (associating dscp values to priorities)
# This example is using two DCBX traffic class priorities to better demonstrate
DCQCN in operation
debugfs.sh -n ens6f0 -t dscp_pfc_enable 1
debugfs.sh -n ens6f0 -t dscp_pfc_set 20 5
debugfs.sh -n ens6f0 -t dscp_pfc_set 22 6

# static DCB configurations.0x10 is static mode.Mark priorities 5 and 6 as
# subject to pfc
debugfs.sh -n ens6f0 -t dcbx_set_mode 0x10
debugfs.sh -n ens6f0 -t dcbx_set_pfc 5 1
debugfs.sh -n ens6f0 -t dcbx_set_pfc 6 1

# set roce global overrides for qp params. enable exn and open QPs with dscp 20
debugfs.sh -n ens6f0 -t rdma_glob_ecn 1
debugfs.sh -n ens6f0 -t rdma_glob_dscp 20

# open some QPs (DSCP 20)
ib_write_bw -d qedr0 -q 16 -F -x 1 --run_infinitely

# change global dscp qp params
debugfs.sh -n ens6f0 -t rdma_glob_dscp 22

# open some more QPs (DSCP 22)
ib_write_bw -d qedr0 -q 16 -F -x 1 -p 8000 --run_infinitely

# observe PFCs being generated on multiple priorities
debugfs.sh -n ens6f0 -d phy_mac_stat -P 0 | grep "Class Based Flow Control"
```

限制

DCQCN 具有以下限制：

- 当前的 DCQCN 模式最多仅支持 64 个 QP。
- Marvell 适配器可从 vLAN 优先级或 ToS 字段中的 DSCP 字节确定用于 PFC 的 vLAN 优先级。但在两者同时存在的情况下，vLAN 优先。

8 iWARP 配置

互联网广域 RDMA 协议 (iWARP) 是一种实现 RDMA 的计算机联网协议，用于通过 IP 网络进行有效的数据传输。 iWARP 设计用于多种环境，包括 LAN、存储网络、数据中心网络和 WAN。

本章提供以下内容的说明：

- 为 iWARP 准备适配器
- 第 175 页上 “ 在 Windows 上配置 iWARP”
- 第 179 页上 “ 在 Linux 上配置 iWARP”

注

某些 iWARP 功能在当前版本中可能并未完全启用。有关详细信息，请参阅附录 D 功能约束。

为 iWARP 准备适配器

本节提供有关使用 HII 预引导适配器 iWARP 配置的说明。有关预引导适配器配置的更多信息，请参阅 第 5 章 适配器预引导配置。

在默认模式下通过 HII 配置 iWARP:

1. 访问服务器的 BIOS System Setup （BIOS 系统设置），然后单击 **Device Settings** （设备设置）。
2. 在 Device Settings （设备设置）页面上，选择用于 25G 41*xxx* 系列适配器的端口。
3. 在所选适配器的 Main Configuration Page （主要配置页面）上，单击 **NIC Configuration** （NIC 配置）。
4. 在 NIC Configuration （NIC 配置）页面上：
   a. 将 **NIC + RDMA Mode** （NIC + RDMA 模式）设置为 **Enabled** （已启用）。
   b. 将 **RDMA Protocol Support** （RDMA 协议支持）设置为 **RoCE/iWARP** 或 **iWARP**。

c. 单击 **Back** （后退）。

5. 在 Main Configuration Page （主要配置页面）上，单击 **Finish** （完成）。
6. 在 Warning - Saving Changes （警告 - 保存更改）消息框中，单击 **Yes** （是）保存配置。
7. 在 Success - Saving Changes （成功 - 保存更改）消息框中，单击 **OK** （确定）。
8. 重复步骤 2 至步骤 7 配置其他端口的 NIC 和 iWARP。
9. 要完成两个端口的适配器准备：
   a. 在 Device Settings （设备设置）页面上，单击 **Finish** （完成）。
   b. 在主菜单上，单击 **Finish** （完成）。
   c. 退出以重新引导系统。

继续 第 175 页上 “ 在 Windows 上配置 iWARP” 或 第 179 页上 “ 在 Linux 上配置 iWARP”。

在 Windows 上配置 iWARP

本节提供启用 iWARP、验证 RDMA 以及验证 Windows 上的 iWARP 流量的程序。有关支持 iWARP 的 OS 列表，请参阅 第 126 页上 表 7-1。

要在 Windows 主机上启用 iWARP 并验证 RDMA：

1. 在 Windows 主机上启用 iWARP。
   a. 打开 Windows 设备管理器，然后打开 41*xxx* 系列适配器 NDIS 微型端口属性。
   b. 在 FastLinQ Adapter Properties （FastLinQ 适配器属性）上，单击 **Advanced** （高级）选项卡。
   c. 在 Advanced （高级）页面的 **Property** （属性）下，请执行以下操作：
      - 选择 **Network Direct Functionality** （Network Direct 功能），然后为 **Value** （值）选择 **Enabled** （已启用）。
      - 选择 **NetworkDirect Technology** （Network Direct 技术），然后为 **Value** （值）选择 **iWARP**。
   d. 单击 **OK** （确定）保存您的更改并关闭适配器属性。

2. 使用 Windows PowerShell，验证 RDMA 是否已启用。
`Get-NetAdapterRdma` 命令输出 （图 8-1）显示支持 RDMA 的适配器。

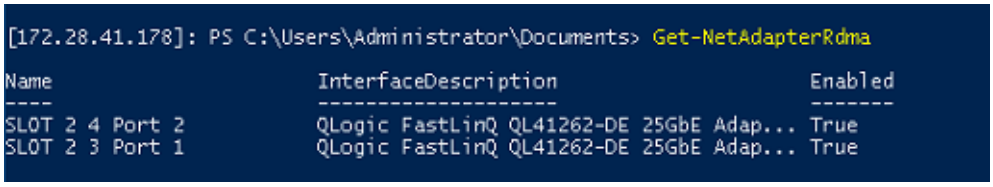


图 8-1. Windows PowerShell 命令：Get-NetAdapterRdma

3. 使用 Windows PowerShell，验证 `NetworkDirect` 是否已启用。
`Get-NetOffloadGlobalSetting` 命令输出 (图 8-2) 将 `NetworkDirect` 显示为 `Enabled` （已启用）。

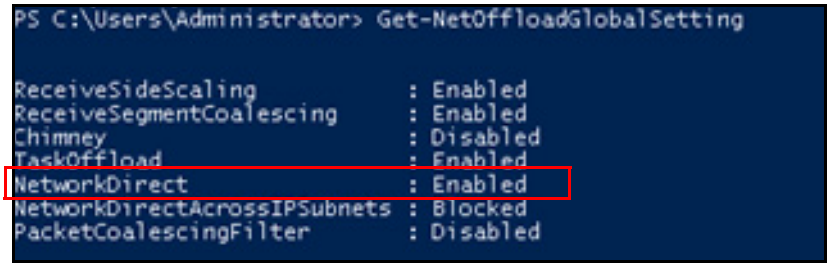


图 8-2. Windows PowerShell 命令：Get-NetOffloadGlobalSetting

要验证 iWARP 流量：

1. 映射 SMB 驱动器，然后运行 iWARP 流量。
2. 启动性能监视器 (Perfmon)。
3. 在 Add Counters （添加计数器）对话框中，单击 **RDMA Activity** （RDMA 活动），然后选择适配器实例。

图 8-3 显示一个示例。

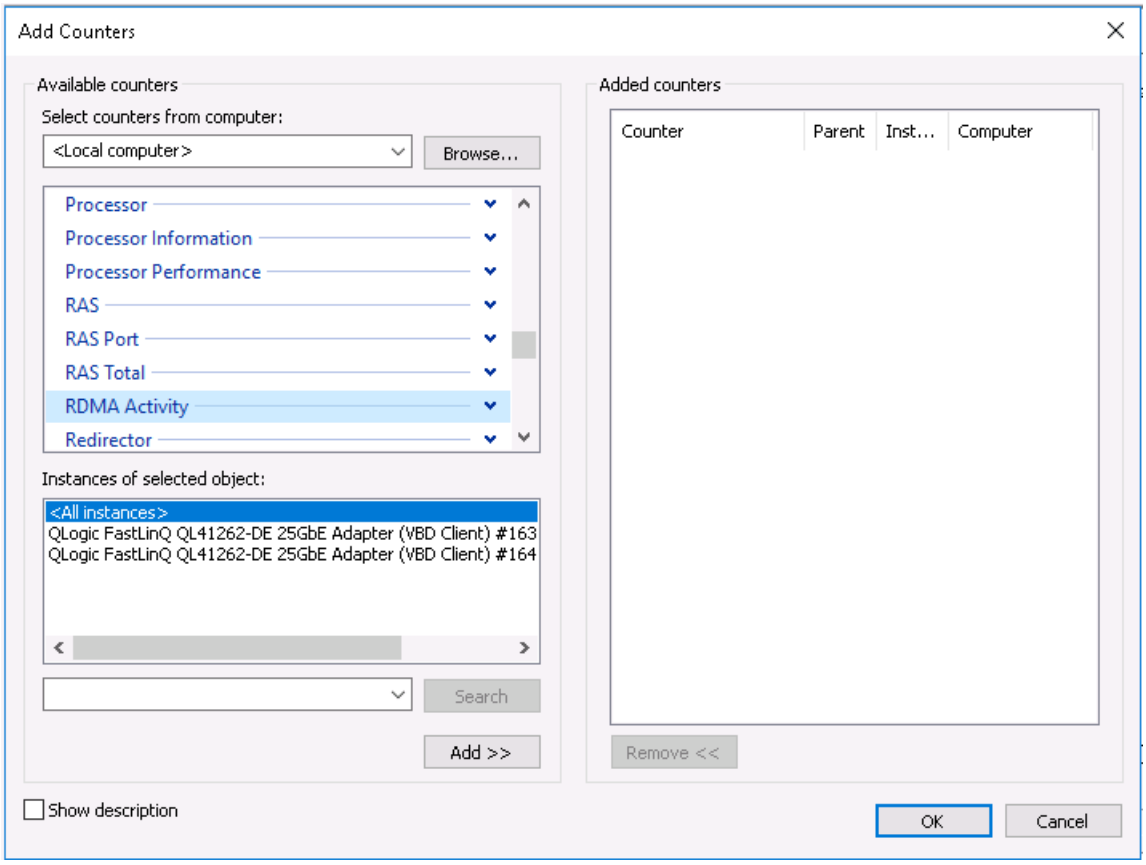


图 8-3. Perfmon：添加计数器

如果 iWARP 流量正在运行，计数器将显示为如图 8-4 示例中所示。

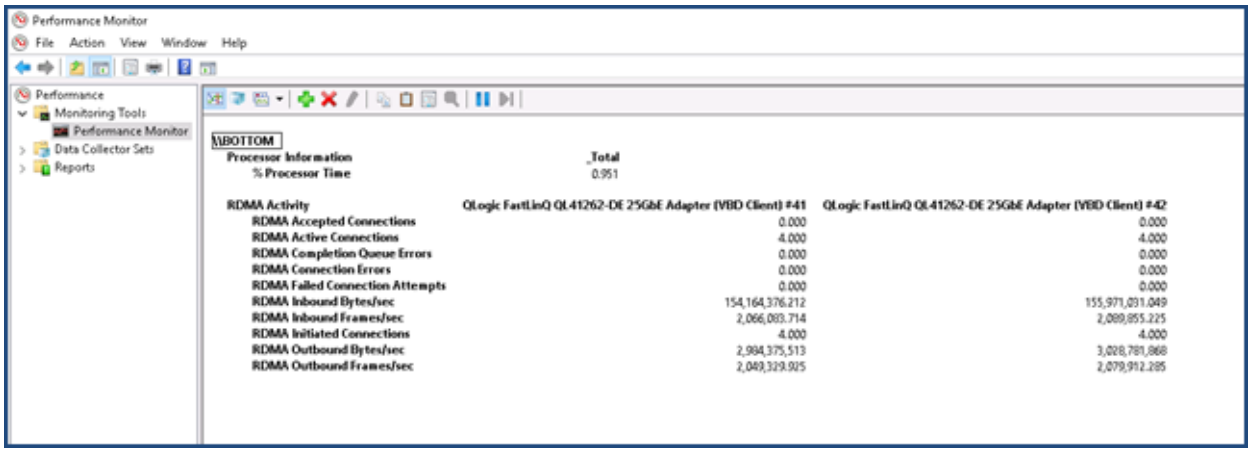


图 8-4. Perfmon：验证 iWARP 流量

注

有关如何在 Windows 中查看 Marvell RDMA 计数器的更多信息，请参阅 第 135 页上 “ 查看 RDMA 计数器 ”。

4. 要验证 SMB 连接：

   a. 在命令提示符处，请如下发出 `net use` 命令：

```
C:\Users\Administrator> net use
New connections will be remembered.

Status    Local     Remote                    Network
-------------------------------------------------------------
OK        F:        \\192.168.10.10\Share1    Microsoft Windows Network
The command completed successfully.
```

   b. 请如下发出 `netstat -xan` 命令，其中 `Share1` 映射为 SMB 共享：

```
C:\Users\Administrator> netstat -xan
Active NetworkDirect Connections, Listeners, ShareEndpoints

 Mode   IfIndex Type       Local Address       Foreign Address      PID

 Kernel      56 Connection 192.168.11.20:16159 192.168.11.10:445    0
 Kernel      56 Connection 192.168.11.20:15903 192.168.11.10:445    0
 Kernel      56 Connection 192.168.11.20:16159 192.168.11.10:445    0
 Kernel      56 Connection 192.168.11.20:15903 192.168.11.10:445    0
```

```
Kernel      60 Listener   [fe80::e11d:9ab5:a47d:4f0a%56]:445 NA    0
Kernel      60 Listener   192.168.11.20:445   NA                  0
Kernel      60 Listener   [fe80::71ea:bdd2:ae41:b95f%60]:445 NA    0
Kernel      60 Listener   192.168.11.20:16159 192.168.11.10:445   0
```

在 Linux 上配置 iWARP

Marvell 41*xxx* 系列适配器 在第 126 页上 表 7-1 中列出的 Linux 开放结构企业分布 (OFED) 上支持 iWARP。

Linux 系统中的 iWARP 配置包括以下各项：

- 安装驱动程序
- 配置 iWARP 和 RoCE
- 检测设备
- 支持的 iWARP 应用程序
- 为 iWARP 运行 Perftest
- 配置 NFS-RDMA

安装驱动程序

如第 3 章 驱动程序安装中所示，安装 RDMA 驱动程序。

配置 iWARP 和 RoCE

注

仅当您在使用 HII 将 **iWARP+RoCE** 选为预引导配置期间的 RDMA Protocol Support （RDMA 协议支持）参数值时，此程序才适用 （请参阅配置 NIC 参数，第 50 页上的步骤 5）。

要启用 iWARP 和 RoCE：

1. 卸载所有 FastLinQ 驱动程序，如下所示：

   ```
   # modprobe -r qedr or modprobe -r qede
   ```

2. 使用以下命令语法，通过使用端口接口 PCI ID (`xx:xx.x`) 和 RDMA 协议值 (*p*) 加载 `qed` 驱动程序来更改 RDMA 协议。

   ```
   # modprobe -v qed rdma_protocol_map=<xx:xx.x-p>
   ```

RDMA 协议 (p) 值如下：

- 0— 接受默认值 (RoCE)
- 1— 无 RDMA
- 2—RoCE
- 3—iWARP

例如，要将 04:00.0 所给端口上的接口从 RoCE 更改为 iWARP，请发出以下命令：

```
# modprobe -v qed rdma_protocol_map=04:00.0-3
```

3. 发出以下命令加载 RDMA 驱动程序：

```
# modprobe -v qedr
```

以下示例显示在多个 NPAR 接口上将 RDMA 协议更改为 iWARP 的命令条目：

```
# modprobe qed rdma_protocol_map=04:00.1-3,04:00.3-3,04:00.5-3,
04:00.7-3,04:01.1-3,04:01.3-3,04:01.5-3,04:01.7-3
# modprobe -v qedr
# ibv_devinfo |grep iWARP
        transport:                      iWARP (1)
        transport:                      iWARP (1)
        transport:                      iWARP (1)
        transport:                      iWARP (1)
        transport:                      iWARP (1)
        transport:                      iWARP (1)
        transport:                      iWARP (1)
        transport:                      iWARP (1)
```

检测设备

要检测设备：

1. 要验证是否检测到 RDMA 设备，请查看 dmesg 日志：

```
# dmesg |grep qedr
[10500.191047] qedr 0000:04:00.0: registered qedr0
[10500.221726] qedr 0000:04:00.1: registered qedr1
```

2. 发出 ibv_devinfo 命令，然后验证传输类型。

如果该命令成功，则每个 PCI 功能将显示单独的 hca_id。例如 （如果检查上述双端口适配器的第二个端口）：

```
[root@localhost ~]# ibv_devinfo -d qedr1
hca_id: qedr1
        transport:                      iWARP (1)
```

```
fw_ver:                         8.14.7.0
node_guid:                      020e:1eff:fec4:c06e
sys_image_guid:                 020e:1eff:fec4:c06e
vendor_id:                      0x1077
vendor_part_id:                 5718
hw_ver:                         0x0
phys_port_cnt:                  1
        port:   1
                state:                  PORT_ACTIVE (4)
                max_mtu:                4096 (5)
                active_mtu:             1024 (3)
                sm_lid:                 0
                port_lid:               0
                port_lmc:               0x00
                link_layer:             Ethernet
```

支持的 iWARP 应用程序

适用于 iWARP 的 Linux 支持的 RDMA 应用程序包括以下各项：

- ibv_devinfo、 ib_devices
- ib_send_bw/lat、 ib_write_bw/lat、 ib_read_bw/lat、 ib_atomic_bw/lat
  对于 iWARP，所有应用程序必须通过 `-R` 选项使用 RDMA 通信管理器 (rdma_cm)。
- rdma_server、 rdma_client
- rdma_xserver、 rdma_xclient
- rping
- RDMA 上的 NFS (NFSoRDMA)
- iSER （有关详细信息，请参阅 第 9 章 iSER 配置）
- NVMe-oF（有关详细信息，请参阅 第 13 章 使用 RDMA 的 NVMe-oF 配置）

为 iWARP 运行 Perftest

所有 perftest 工具均通过 iWARP 传输类型受支持。您必须使用 RDMA 连接管理器（通过 `-R` 选项）运行工具。

示例：

1. 在一个服务器上，发出以下命令 （在此示例中使用第二个端口）：

```
# ib_send_bw -d qedr1 -F -R
```

2. 在一个客户端上，发出以下命令（在此示例中使用第二个端口）：

```
[root@localhost ~]# ib_send_bw -d qedr1 -F -R 192.168.11.3
---------------------------------------------------------------------------------
                    Send BW Test
 Dual-port       : OFF          Device         : qedr1
 Number of qps   : 1            Transport type : IW
 Connection type : RC           Using SRQ      : OFF
 TX depth        : 128
 CQ Moderation   : 100
 Mtu             : 1024[B]
 Link type       : Ethernet
 GID index       : 0
 Max inline data : 0[B]
 rdma_cm QPs     : ON
 Data ex. method : rdma_cm
---------------------------------------------------------------------------------
 local address: LID 0000 QPN 0x0192 PSN 0xcde932
 GID: 00:14:30:196:192:110:00:00:00:00:00:00:00:00:00:00
 remote address: LID 0000 QPN 0x0098 PSN 0x46fffc
 GID: 00:14:30:196:195:62:00:00:00:00:00:00:00:00:00:00
---------------------------------------------------------------------------------
 #bytes     #iterations    BW peak[MB/sec]    BW average[MB/sec]   MsgRate[Mpps]
 65536      1000             2250.38            2250.36              0.036006
---------------------------------------------------------------------------------
```

注

对于延迟应用程序（发送 / 写入），如果 perftest 版本为最新（例如，perftest-3.0-0.21.g21dc344.x86_64.rpm），请使用支持的内嵌大小值：0-128。

配置 NFS-RDMA

适用于 iWARP 的 NFS-RDMA 包括服务器和客户端配置步骤。

要配置 NFS 服务器：

1. 通过发出以下命令创建 nfs-server 目录并授予权限：

```
# mkdir /tmp/nfs-server
# chmod 777 /tmp/nfs-server
```

2. 在 `/etc/exports` 文件中，对您必须使用服务器上 NFS-RDMA 导出的目录，创建以下条目：

   /tmp/nfs-server *(rw,fsid=0,async,insecure,no_root_squash)

   确保对您导出的每个目录使用不同的文件系统标识 (FSID)。

3. 请如下加载 svcrdma 模块：

   # **modprobe svcrdma**

4. 加载服务如下：

   - 对于 SLES，启用并启动 NFS 服务器别名：

     # **systemctl enable|start|status nfsserver**

   - 对于 RHEL，启用并启动 NFS 服务器和服务：

     # **systemctl enable|start|status nfs**

5. 请如下在此文件中包括默认 RDMA 端口 20049：

   # **echo rdma 20049 > /proc/fs/nfsd/portlist**

6. 要使本地目录可供 NFS 客户端进行装载，请如下发出 `exportfs` 命令：

   # **exportfs -v**

要配置 NFS 客户端：

注

此 NFS 客户端配置程序也适用于 RoCE。

1. 通过发出以下命令创建 `nfs-client` 目录并授予权限：

   # **mkdir /tmp/nfs-client**
   # **chmod 777 /tmp/nfs-client**

2. 请如下加载 xprtrdma 模块：

   # **modprobe xprtrdma**

3. 根据您的版本，装载 NFS 文件系统：

   对于 NFS 版本 3：

   # **mount -o rdma,port=20049 192.168.2.4:/tmp/nfs-server /tmp/nfs-client**

对于 NFS 版本 4：

```
# mount -t nfs4 -o rdma,port=20049 192.168.2.4:/tmp/nfs-server
/tmp/nfs-client
```

注

NFSoRDMA 的默认端口为 20049。但是，与 NFS 客户端对齐的任何其他端口也将起工作。

4. 通过发出 `mount` 命令，验证文件系统已装载。确保 RDMA 端口和文件系统版本正确无误。

```
# mount |grep rdma
```

9 iSER 配置

本章提供为 Linux （RHEL 和 SLES）和 VMware ESXi 6.7 配置 RDMA 的 iSCSI 扩展 (iSER) 的步骤，包括：

准备工作

在准备配置 iSER 时，请考虑以下事项：

- 仅在以下操作系统的内建 OFED 中支持 iSER：
  - RHEL 8.*x*
  - RHEL 7.6 及更高版本
  - SLES 12 SP4 及更高版本
  - SLES 15 SP0 及更高版本
  - VMware ESXi 6.7 U1
- 在登录目标之后或运行 I/O 流量时，卸载 Linux RoCE qedr 驱动程序可能导致系统崩溃。
- 在运行 I/O 时，执行接口关闭 / 上线测试或执行电缆拉力测试可能会导致驱动程序或 iSER 模块错误，而这些错误可能导致系统崩溃。如果发生这种情况，请重新引导系统。

为 RHEL 配置 iSER

要为 RHEL 配置 iSER：

1. 如第 149 页上 “RHEL 的 RoCE 配置 ”中所述，安装内建 OFED。

   注

   iSER 不支持非内建 OFED，因为 ib_isert 模块在非内建 OFED 3.18-2 GA/3.18-3 GA 版本中不可用。内建 ib_isert 模块不能与任何非内建 OFED 版本一起使用。

2. 如第 10 页上 “ 移除 Linux 驱动程序 ”中所述，卸载任何现有 FastLinQ 驱动程序。
3. 如第 14 页上 “ 安装包含 RDMA 的 Linux 驱动程序 ”中所述，安装最新 FastLinQ 驱动程序和 libqedr 软件包。
4. 如下所示加载 RDMA 服务；

   ```
   systemctl start rdma
   modprobe qedr
   modprobe ib_iser
   modprobe ib_isert
   ```

5. 通过发出 `lsmod | grep qed` 和 `lsmod | grep iser` 命令，验证在启动器和目标设备上加载的所有 RDMA 和 iSER 模块。
6. 如第 152 页上的步骤 6中所示，通过发出 `ibv_devinfo` 命令，验证存在独立的 `hca_id` 实例。
7. 检查启动器设备和目标设备上的 RDMA 连接。

   a. 在启动器设备上，发出以下命令：

   ```
   rping -s -C 10 -v
   ```

   b. 在目标设备上，发出以下命令：

   ```
   rping -c -a 192.168.100.99 -C 10 -v
   ```

图 9-1 显示了成功 RDMA ping 操作的示例。

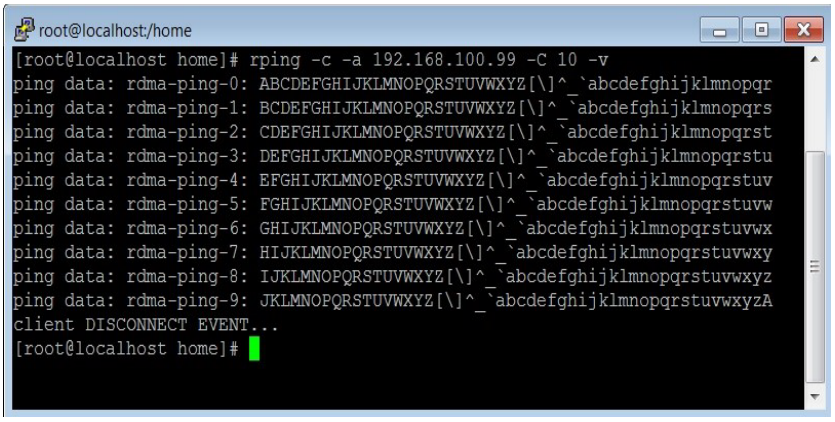


图 9-1. RDMA Ping 操作成功

8. 可以使用 Linux TCM-LIO 目标来测试 iSER。其设置与任何 iSCSI 目标相同，但要在适用门户上发出命令 `enable_iser Boolean=true`。门户实例在图 9-2 中标识为 **`iser`**。

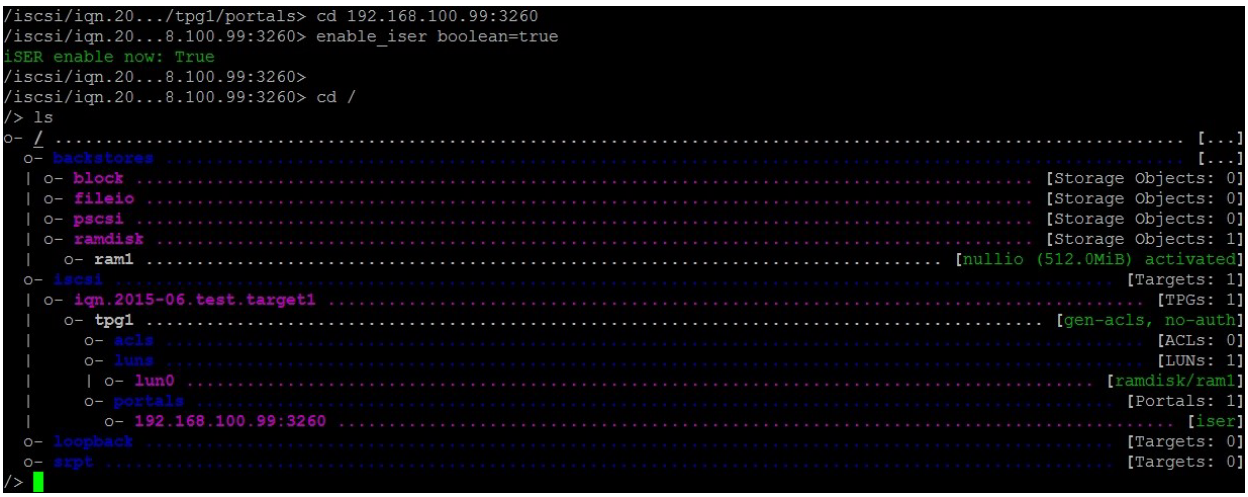


图 9-2. iSER 门户实例

9. 使用 `yum install iscsi-initiator-utils` 命令安装 Linux iSCSI 启动器公用程序。

   a. 要查找 iSER 目标，请发出 `iscsiadm` 命令。例如：

   `iscsiadm -m discovery -t st -p 192.168.100.99:3260`

b. 要将传输模式更改为 iSER，请发出 `iscsiadm` 命令。例如：

```
iscsiadm -m node -T iqn.2015-06.test.target1 -o update -n
iface.transport_name -v iser
```

c. 要连接到或登录到 iSER 目标，请发出 `iscsiadm` 命令。例如：

```
iscsiadm -m node -l -p 192.168.100.99:3260 -T
iqn.2015-06.test.target1
```

d. 确认目标连接中的 `Iface Transport` 为 **`iser`**，如图 9-3 中所示。发出 `iscsiadm` 命令；例如：

```
iscsiadm -m session -P2
```

```
[root@localhost ~]# iscsiadm -m discovery -t st -p 192.168.100.99:3260
192.168.100.99:3260,1 iqn.2015-06.test.target1
192.168.100.99:3260,1 iqn.2015-06.test.target1
[root@localhost ~]#
[root@localhost ~]# iscsiadm -m node -T iqn.2015-06.test.target1 -o update -n iface.transport_name -v iser
[root@localhost ~]#
[root@localhost ~]#
[root@localhost ~]# iscsiadm -m node -l -p 192.168.100.99:3260 -T iqn.2015-06.test.target1
Logging in to [iface: default, target: iqn.2015-06.test.target1, portal: 192.168.100.99,3260] (multiple)
Login to [iface: default, target: iqn.2015-06.test.target1, portal: 192.168.100.99,3260] successful.
[root@localhost ~]#
[root@localhost ~]# iscsiadm -m session -P2
Target: iqn.2015-06.test.target1 (non-flash)
        Current Portal: 192.168.100.99:3260,1
        Persistent Portal: 192.168.100.99:3260,1
                **********
                Interface:
                **********
                Iface Name: default
                Iface Transport: iser
                Iface Initiatorname: iqn.1994-05.com.redhat:c672dfb8b08f
                Iface IPaddress: <empty>
                Iface HWaddress: <empty>
                Iface Netdev: <empty>
                SID: 33
                iSCSI Connection State: LOGGED IN
                iSCSI Session State: LOGGED_IN
                Internal iscsid Session State: NO CHANGE
                *********
                Timeouts:
                *********
                Recovery Timeout: 120
```

图 9-3. Iface 传输确认

e. 要检查新 iSCSI 设备，如图 9-4 中所示，发出 `lsscsi` 命令。

```
[root@localhost ~]# lsscsi
[6:0:0:0]    disk    HP       LOGICAL VOLUME   1.18  /dev/sdb
[6:0:0:1]    disk    HP       LOGICAL VOLUME   1.18  /dev/sda
[6:0:0:3]    disk    HP       LOGICAL VOLUME   1.18  /dev/sdc
[6:3:0:0]    storage HP       P440ar           1.18  -
[39:0:0:0]   disk    LIO-ORG  ram1             4.0   /dev/sdd
[root@localhost ~]#
```

图 9-4. 检查新 iSCSI 设备

为 SLES 12 及更高版本配置 iSER

由于 targetcli 没有在 SLES 12 及更高版本上内建，您必须完成以下步骤。

要为 SLES 12 及更高版本配置 iSER：

1. 安装 targetcli。

   对于 SLES 12：

   从 ISO 映像 （x86_64 和 noarch 位置）定位、复制并安装以下 RPM：

   ```
   lio-utils-4.1-14.6.x86_64.rpm
   python-configobj-4.7.2-18.10.noarch.rpm
   python-PrettyTable-0.7.2-8.5.noarch.rpm
   python-configshell-1.5-1.44.noarch.rpm
   python-pyparsing-2.0.1-4.10.noarch.rpm
   python-netifaces-0.8-6.55.x86_64.rpm
   python-rtslib-2.2-6.6.noarch.rpm
   python-urwid-1.1.1-6.144.x86_64.rpm
   targetcli-2.1-3.8.x86_64.rpm
   ```

   对于 SLES 15 和 SLES 15 SP1：

   通过发出以下 Zypper 命令加载 SLES 包 DVD 并安装 targetcli，该命令会将安装所有依赖包：

   ```
   # zypper install python3-targetcli-fb
   ```

2. 启动 targetcli 之前，请如下加载所有 RoCE 设备驱动程序和 iSER 模块：

   ```
   # modprobe qed
   # modprobe qede
   # modprobe qedr
   # modprobe ib_iser    （启动器）
   # modprobe ib_isert   （目标）
   ```

3. 配置 iSER 目标之前，配置 NIC 接口并运行 L2 和 RoCE 流量，如第 152 页 上的步骤 7 中所述。

4. 对于 SLES 15 和 SLES 15 SP1，插入 SLES 包 DVD 并安装 targetcli 公用程序。此命令还会安装所有依赖包。

   ```
   # zypper install python3-targetcli-fb
   ```

5. 启动 targetcli 公用程序，然后在 iSER 目标系统上配置您的目标。

注

RHEL 和 SLES 中的 targetcli 版本是不同的。请务必使用合适的后备存储配置您的目标：

- RHEL 使用 *ramdisk*
- SLES 使用 *rd_mcp*

在 RHEL 和 SLES 上通过 iWARP 使用 iSER

配置与 RoCE 类似的 iSER 启动器和目标以与 iWARP 一起使用。您可以使用不同的方法创建 Linux-IO (LIO™) 目标；本节中列出了一种方法。在 SLES 12 和 RHEL 7.*x* 中的 targetcli 配置可能会由于版本不同而存在一些差异。

要为 LIO 配置目标：

1. 使用 targetcli 公用程序创建 LIO 目标。发出以下命令：

```
# targetcli
targetcli shell version 2.1.fb41
Copyright 2011-2013 by Datera, Inc and others.
For help on commands, type 'help'.
```

2. 发出以下命令：

```
/> /backstores/ramdisk create Ramdisk1-1 1g nullio=true
/> /iscsi create iqn.2017-04.com.org.iserport1.target1
/> /iscsi/iqn.2017-04.com.org.iserport1.target1/tpg1/luns create /backstores/ramdisk/Ramdisk1-1
/> /iscsi/iqn.2017-04.com.org.iserport1.target1/tpg1/portals/ create 192.168.21.4 ip_port=3261
/> /iscsi/iqn.2017-04.com.org.iserport1.target1/tpg1/portals/192.168.21.4:3261 enable_iser
boolean=true
/> /iscsi/iqn.2017-04.com.org.iserport1.target1/tpg1 set attribute authentication=0
demo_mode_write_protect=0 generate_node_acls=1 cache_dynamic_acls=1
/> saveconfig
```

图 9-5 显示 LIO 的目标配置。

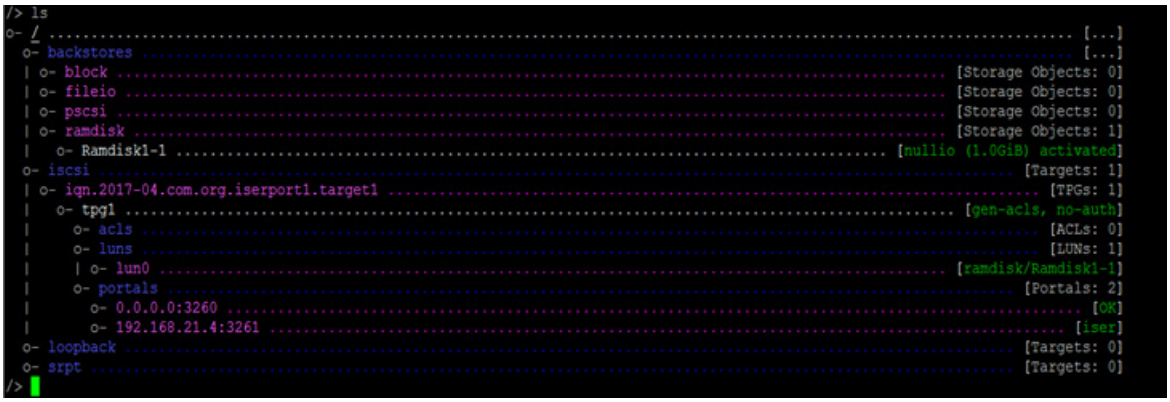


图 9-5. LIO 目标配置

要为 iWARP 配置启动器：

1. 要使用端口 3261 查找 iSER LIO 目标，请如下发出 `iscsiadm` 命令：

```
# iscsiadm -m discovery -t st -p 192.168.21.4:3261 -I iser
192.168.21.4:3261,1 iqn.2017-04.com.org.iserport1.target1
```

2. 将传输模式更改为 `iser`，如下所示：

```
# iscsiadm -m node -o update -T iqn.2017-04.com.org.iserport1.target1 -n
iface.transport_name -v iser
```

3. 使用端口 3261 登录到目标：

```
# iscsiadm -m node -l -p 192.168.21.4:3261 -T
iqn.2017-04.com.org.iserport1.target1
Logging in to [iface: iser, target: iqn.2017-04.com.org.iserport1.target1,
portal: 192.168.21.4,3261] (multiple)
Login to [iface: iser, target: iqn.2017-04.com.org.iserport1.target1, portal:
192.168.21.4,3261] successful.
```

4. 通过发出以下命令，确保这些 LUN 可见：

```
# lsscsi
[1:0:0:0]    storage HP       P440ar           3.56  -
[1:1:0:0]    disk    HP       LOGICAL VOLUME   3.56  /dev/sda
[6:0:0:0]    cd/dvd  hp       DVD-ROM DUD0N    UMD0  /dev/sr0
[7:0:0:0]    disk    LIO-ORG  Ramdisk1-1       4.0   /dev/sdb
```

优化 Linux 性能

考虑本节中介绍的以下 Linux 性能配置增强。

- 将 CPU 配置为最高性能模式
- 配置内核 sysctl 设置
- 配置 IRQ 关联设置
- 配置块设备暂存

将 CPU 配置为最高性能模式

使用以下脚本将 CPU scaling governor 配置为 performance，从而将所有 CPU 设置为最高性能模式：

```
for CPUFREQ in
/sys/devices/system/cpu/cpu*/cpufreq/scaling_governor; do [ -f
$CPUFREQ ] || continue; echo -n performance > $CPUFREQ; done
```

通过发出以下命令，验证所有 CPU 核心是否设置为最高性能模式：

```
cat /sys/devices/system/cpu/cpu*/cpufreq/scaling_governor
```

配置内核 sysctl 设置

按如下所示设定内核 sysctl 设置：

```
sysctl -w net.ipv4.tcp_mem="4194304 4194304 4194304"
sysctl -w net.ipv4.tcp_wmem="4096 65536 4194304"
sysctl -w net.ipv4.tcp_rmem="4096 87380 4194304"
sysctl -w net.core.wmem_max=4194304
sysctl -w net.core.rmem_max=4194304
sysctl -w net.core.wmem_default=4194304
sysctl -w net.core.rmem_default=4194304
sysctl -w net.core.netdev_max_backlog=250000
sysctl -w net.ipv4.tcp_timestamps=0
sysctl -w net.ipv4.tcp_sack=1
sysctl -w net.ipv4.tcp_low_latency=1
sysctl -w net.ipv4.tcp_adv_win_scale=1
echo 0 > /proc/sys/vm/nr_hugepages
```

配置 IRQ 关联设置

以下示例将 CPU 核心 0、1、2 和 3 分别设置为中断请求 (IRQ) XX、YY、ZZ 和 XYZ。对分配给端口的各 IRQ 执行以下步骤 （默认为每端口八个队列）。

```
systemctl disable irqbalance
systemctl stop irqbalance
cat /proc/interrupts | grep qedr 显示分配给每个端口队列的 IRQ
echo 1 > /proc/irq/XX/smp_affinity_list
echo 2 > /proc/irq/YY/smp_affinity_list
echo 4 > /proc/irq/ZZ/smp_affinity_list
echo 8 > /proc/irq/XYZ/smp_affinity_list
```

配置块设备暂存

请如下设定每个 iSCSI 设备或目标的块设备暂存设置：

```
echo noop > /sys/block/sdd/queue/scheduler
echo 2 > /sys/block/sdd/queue/nomerges
echo 0 > /sys/block/sdd/queue/add_random
echo 1 > /sys/block/sdd/queue/rq_affinity
```

在 ESXi 6.7 上配置 iSER

本节介绍为 VMware ESXi 6.7 配置 iSER 的有关信息。

准备工作

为 ESXi 6.7 配置 iSER 前，请确保完成以下操作：

- 在 ESXi 6.7 系统上安装带 NIC 和 RoCE 驱动器的 CNA 程序包并列出设备。为了查看 RDMA 设备，请发出以下命令：

```
esxcli rdma device list
Name     Driver   State    MTU  Speed    Paired Uplink  Description
-------  -------  ------  ----  -------  -------------  ------------------------------------
vmrdma0  qedrntv  Active  1024  40 Gbps  vmnic4         QLogic FastLinQ QL45xxx RDMA Interface
vmrdma1  qedrntv  Active  1024  40 Gbps  vmnic5         QLogic FastLinQ QL45xxx RDMA Interface
[root@localhost:~] esxcfg-vmknic -l
Interface  Port Group/DVPort/Opaque Network        IP Family IP Address
Netmask         Broadcast       MAC Address       MTU     TSO MSS   Enabled Type
NetStack
vmk0       Management Network                     IPv4      172.28.12.94
255.255.240.0   172.28.15.255   e0:db:55:0c:5f:94 1500    65535     true    DHCP
defaultTcpipStack
```

```
vmk0       Management Network                    IPv6      fe80::e2db:55ff:fe0c:5f94
64                                   e0:db:55:0c:5f:94 1500    65535     true    STATIC, PREFERRED
defaultTcpipStack
```

- 配置 iSER 目标，从而实现与 iSER 启动器的通信。

为 ESXi 6.7 配置 iSER

为 ESXi 6.7 配置 iSER 的步骤：

1. 通过发出以下命令添加 iSER 设备：

```
esxcli rdma iser add
esxcli iscsi adapter list
Adapter  Driver  State    UID            Description
-------  ------  -------  -------------  -----------------------------------
vmhba64  iser    unbound  iscsi.vmhba64  VMware iSCSI over RDMA (iSER) Adapter
vmhba65  iser    unbound  iscsi.vmhba65  VMware iSCSI over RDMA (iSER) Adapter
```

2. 如下所示禁用防火墙。

```
esxcli network firewall set --enabled=false
esxcli network firewall unload
vsish -e set /system/modules/iscsi_trans/loglevels/iscsitrans 0
vsish -e set /system/modules/iser/loglevels/debug 4
```

3. 创建标准 vSwitch VMkernel 端口组并分配 IP：

```
esxcli network vswitch standard add -v vSwitch_iser1
esxcfg-nics -l
Name    PCI          Driver      Link Speed      Duplex MAC Address       MTU    Description
vmnic0  0000:01:00.0 ntg3        Up   1000Mbps   Full   e0:db:55:0c:5f:94 1500   Broadcom
Corporation NetXtreme BCM5720 Gigabit Ethernet
vmnic1  0000:01:00.1 ntg3        Down 0Mbps      Half   e0:db:55:0c:5f:95 1500   Broadcom
Corporation NetXtreme BCM5720 Gigabit Ethernet
vmnic2  0000:02:00.0 ntg3        Down 0Mbps      Half   e0:db:55:0c:5f:96 1500   Broadcom
Corporation NetXtreme BCM5720 Gigabit Ethernet
vmnic3  0000:02:00.1 ntg3        Down 0Mbps      Half   e0:db:55:0c:5f:97 1500   Broadcom
Corporation NetXtreme BCM5720 Gigabit Ethernet
vmnic4  0000:42:00.0 qedentv     Up   40000Mbps  Full   00:0e:1e:d5:f6:a2 1500   QLogic Corp.
QLogic FastLinQ QL41xxx 10/25/40/50/100 GbE Ethernet Adapter
vmnic5  0000:42:00.1 qedentv     Up   40000Mbps  Full   00:0e:1e:d5:f6:a3 1500   QLogic Corp.
QLogic FastLinQ QL41xxx 10/25/40/50/100 GbE Ethernet Adapter

esxcli network vswitch standard uplink add -u vmnic5 -v vSwitch_iser1
esxcli network vswitch standard portgroup add -p "rdma_group1" -v vSwitch_iser1
esxcli network ip interface add -i vmk1 -p "rdma_group1"
esxcli network ip interface ipv4 set -i vmk1 -I 192.168.10.100 -N 255.255.255.0 -t static
esxcfg-vswitch -p "rdma_group1" -v 4095 vSwitch_iser1
```

```
esxcli iscsi networkportal add -A vmhba67 -n vmk1
esxcli iscsi networkportal list
esxcli iscsi adapter get -A vmhba65
vmhba65
   Name: iqn.1998-01.com.vmware:localhost.punelab.qlogic.com qlogic.org qlogic.com
mv.qlogic.com:1846573170:65
   Alias: iser-vmnic5
   Vendor: VMware
   Model: VMware iSCSI over RDMA (iSER) Adapter
   Description: VMware iSCSI over RDMA (iSER) Adapter
   Serial Number: vmnic5
   Hardware Version:
   Asic Version:
   Firmware Version:
   Option Rom Version:
   Driver Name: iser-vmnic5
   Driver Version:
   TCP Protocol Supported: false
   Bidirectional Transfers Supported: false
   Maximum Cdb Length: 64
   Can Be NIC: true
   Is NIC: true
   Is Initiator: true
   Is Target: false
   Using TCP Offload Engine: true
   Using ISCSI Offload Engine: true
```

4. 如下所示将目标添加至 iSER 启动器：

```
esxcli iscsi adapter target list
esxcli iscsi adapter discovery sendtarget add -A vmhba65 -a 192.168.10.11
esxcli iscsi adapter target list
Adapter  Target                   Alias  Discovery Method  Last Error
-------  -----------------------  -----  ----------------  ----------
vmhba65  iqn.2015-06.test.target1         SENDTARGETS       No Error
esxcli storage core adapter rescan --adapter vmhba65
```

5. 如下所示列出所附接的目标：

```
esxcfg-scsidevs -l
mpx.vmhba0:C0:T4:L0
   Device Type: CD-ROM
   Size: 0 MB
   Display Name: Local TSSTcorp CD-ROM (mpx.vmhba0:C0:T4:L0)
   Multipath Plugin: NMP
```

```
   Console Device: /vmfs/devices/cdrom/mpx.vmhba0:C0:T4:L0
   Devfs Path: /vmfs/devices/cdrom/mpx.vmhba0:C0:T4:L0
   Vendor: TSSTcorp  Model: DVD-ROM SN-108BB  Revis: D150
   SCSI Level: 5  Is Pseudo: false Status: on
   Is RDM Capable: false Is Removable: true
   Is Local: true  Is SSD: false
   Other Names:
      vml.0005000000766d686261303a343a30
   VAAI Status: unsupported
naa.6001405e81ae36b771c418b89c85dae0
   Device Type: Direct-Access
   Size: 512 MB
   Display Name: LIO-ORG iSCSI Disk (naa.6001405e81ae36b771c418b89c85dae0)
   Multipath Plugin: NMP
   Console Device: /vmfs/devices/disks/naa.6001405e81ae36b771c418b89c85dae0
   Devfs Path: /vmfs/devices/disks/naa.6001405e81ae36b771c418b89c85dae0
   Vendor: LIO-ORG   Model: ram1             Revis: 4.0
   SCSI Level: 5  Is Pseudo: false Status: degraded
   Is RDM Capable: true  Is Removable: false
   Is Local: false Is SSD: false
   Other Names:
      vml.02000000006001405e81ae36b771c418b89c85dae072616d312020
   VAAI Status: supported
naa.690b11c0159d050018255e2d1d59b612
```

10 iSCSI 配置

本章提供以下 iSCSI 配置信息：

- iSCSI 引导
- 第 197 页上 “Windows Server 中的 iSCSI 卸载 ”
- 第 206 页上 “Linux 环境中的 iSCSI 卸载 ”

注

某些 iSCSI 功能在当前版本中可能并未完全启用。有关详细信息，请参阅 附录 D 功能约束。

要启用 iSCSI 卸载模式，请参阅位于 https://www.marvell.com/documents/5aa5otcbkr0im3ynera3/ 的 *Application Note, Enabling Storage Offloads on Dell and Marvell FastLinQ 41000 Series Adapters*。

iSCSI 引导

Marvell 4*xxxx* 系列千兆位以太网 (GbE) 适配器支持 iSCSI 引导，从而实现无盘系统的操作系统网络引导。iSCSI 引导允许 Windows、Linux 或 VMware 操作系统通过标准 IP 网络从位于远程的 iSCSI 目标计算机引导。

适配器用作 NDIS 或 HBA 卸载装置时，仅在 Windows OS 支持带 iSCSI 引导的巨型帧。

如需从 SAN 的 iSCSI 引导信息，请参阅 第 6 章 从 SAN 引导配置。

Windows Server 中的 iSCSI 卸载

iSCSI 卸载是一种将 iSCSI 协议处理开销从主机处理器卸载到 iSCSI HBA 的技术。iSCSI 卸载可提高网络性能和吞吐量，同时帮助优化服务器处理器的使用。本节介绍如何为 Marvell 41*xxx* 系列适配器配置 Windows iSCSI 卸载功能。

通过适当的 iSCSI 卸载许可，可配置具有 iSCSI 功能的 41*xxx* 系列适配器，以便从主机处理器卸载 iSCSI 处理。以下各节介绍如何使系统能够充分利用 Marvell 的 iSCSI 卸载功能：

- 安装 Marvell 驱动程序
- 安装 Microsoft iSCSI 启动器
- 配置 Microsoft 启动器以使用 Marvell 的 iSCSI 卸载
- iSCSI 卸载常见问题
- Windows Server 2012 R2、 2016 和 2019 iSCSI 引导安装
- iSCSI 故障转储

安装 Marvell 驱动程序

如第 17 页上 “ 安装 Windows 驱动程序软件 ”中所述，安装 Windows 驱动程序。

安装 Microsoft iSCSI 启动器

启动 Microsoft iSCSI 启动器小程序。初次启动时，系统会提示自动服务启动。确认要启动的小程序的选择。

配置 Microsoft 启动器以使用 Marvell 的 iSCSI 卸载

在为 iSCSI 适配器配置 IP 地址后，必须使用 Microsoft 启动器来配置和添加到使用 Marvell FastLinQ iSCSI 适配器的 iSCSI 目标的连接。有关 Microsoft 启动器的更多详细信息，请参阅 Microsoft 用户指南。

要配置 Microsoft 启动器：

1. 打开 Microsoft 启动器。
2. 要根据您的设置配置启动器 IQN 名称，请按照以下步骤操作：
   a. 在 iSCSI Initiator Properties （iSCSI 启动器属性）上，单击 **Configuration** （配置）选项卡。
   b. 在 Configuration （配置）页面 （图 10-1）上，单击 **To modify the initiator name** （修改启动器名称）旁边的 **Change** （更改）。

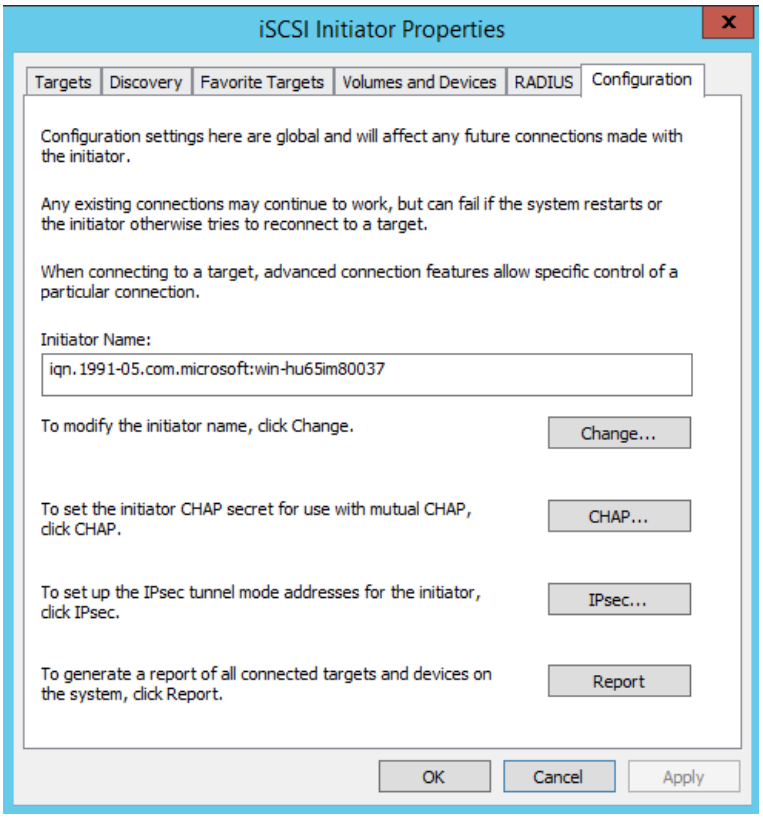


图 10-1. iSCSI 启动器属性，配置页面

c. 在 iSCSI Initiator Name （iSCSI 启动器名称）对话框中，键入新的启动器 IQN 名称，然后单击 **OK** （确定）。（图 10-2）

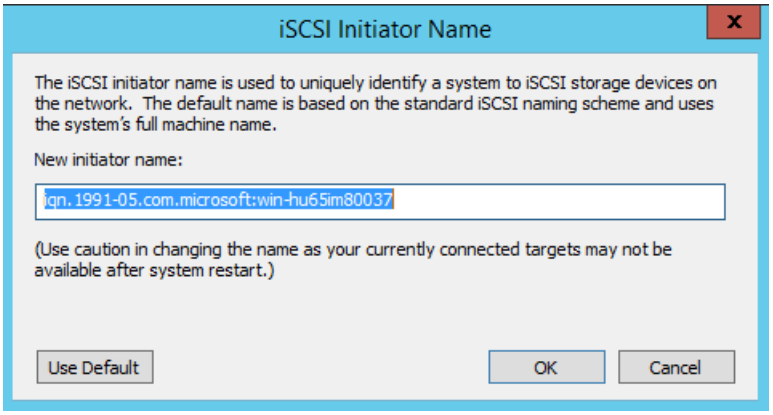


图 10-2. iSCSI 启动器节点名称更改

3. 在 iSCSI Initiator Properties （iSCSI 启动器属性）上，单击 **Discovery** （查找）选项卡。
4. 在 Discovery （查找）页面 （图 10-3）的 **Target portals** （目标门户）下，单击 **Discover Portal** （查找门户）。

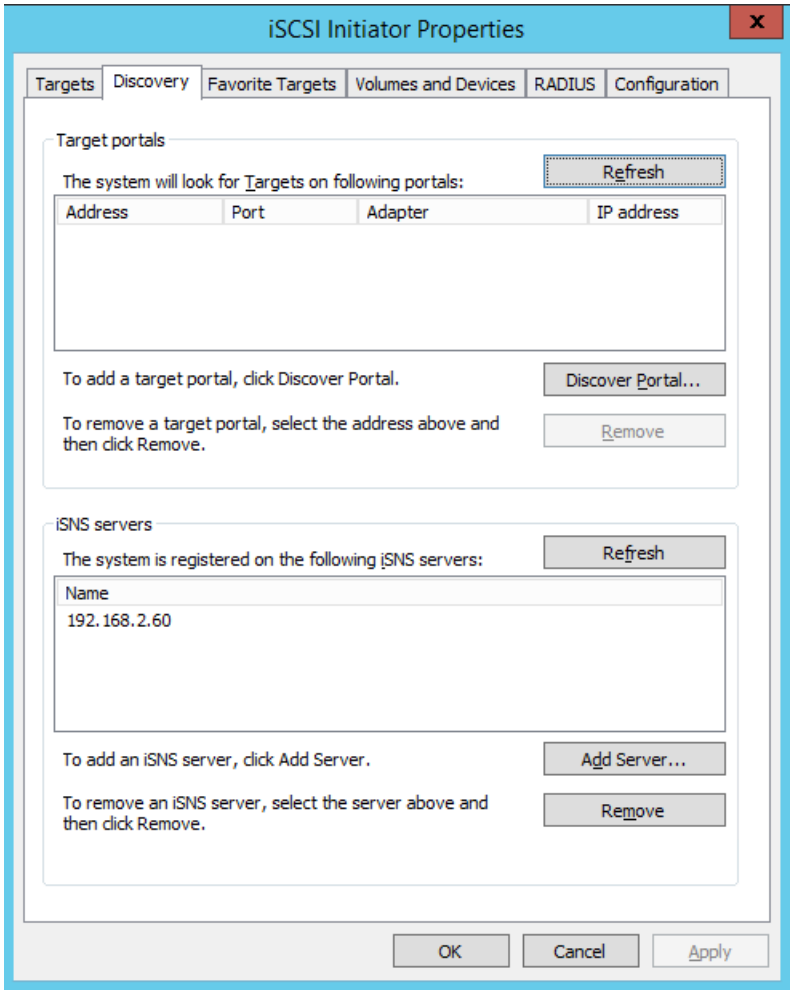


图 10-3. iSCSI 启动器 - 查找目标门户

5. 在 Discover Target Portal （查找目标门户）对话框 （图 10-4）中：
   a. 在 **IP address or DNS name** （IP 地址或 DNS 名称）框中，键入目标的 IP 地址。
   b. 单击 **Advanced** （高级）。

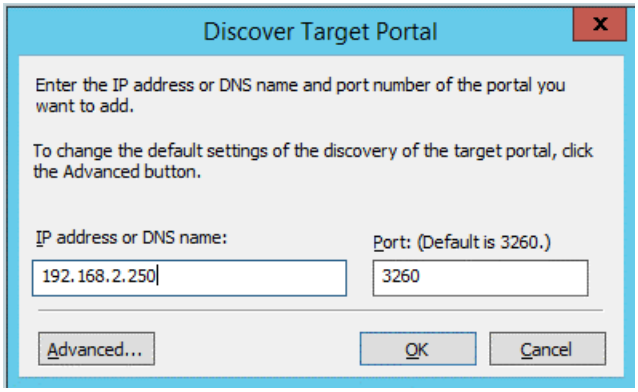


图 10-4. 目标门户 IP 地址

6. 在 Advanced Settings （高级设置）对话框 （图 10-5）中，填写 **Connect using** （连接方式）下的以下各项：
   a. 对于 **Local adapter**（本地适配器），选择 **QLogic <name or model> Adapter** （QLogic < 名称或型号 > 适配器）。
   b. 对于 **Initiator IP** （启动器 IP），选择适配器 IP 地址。
   c. 单击 **OK** （确定）。

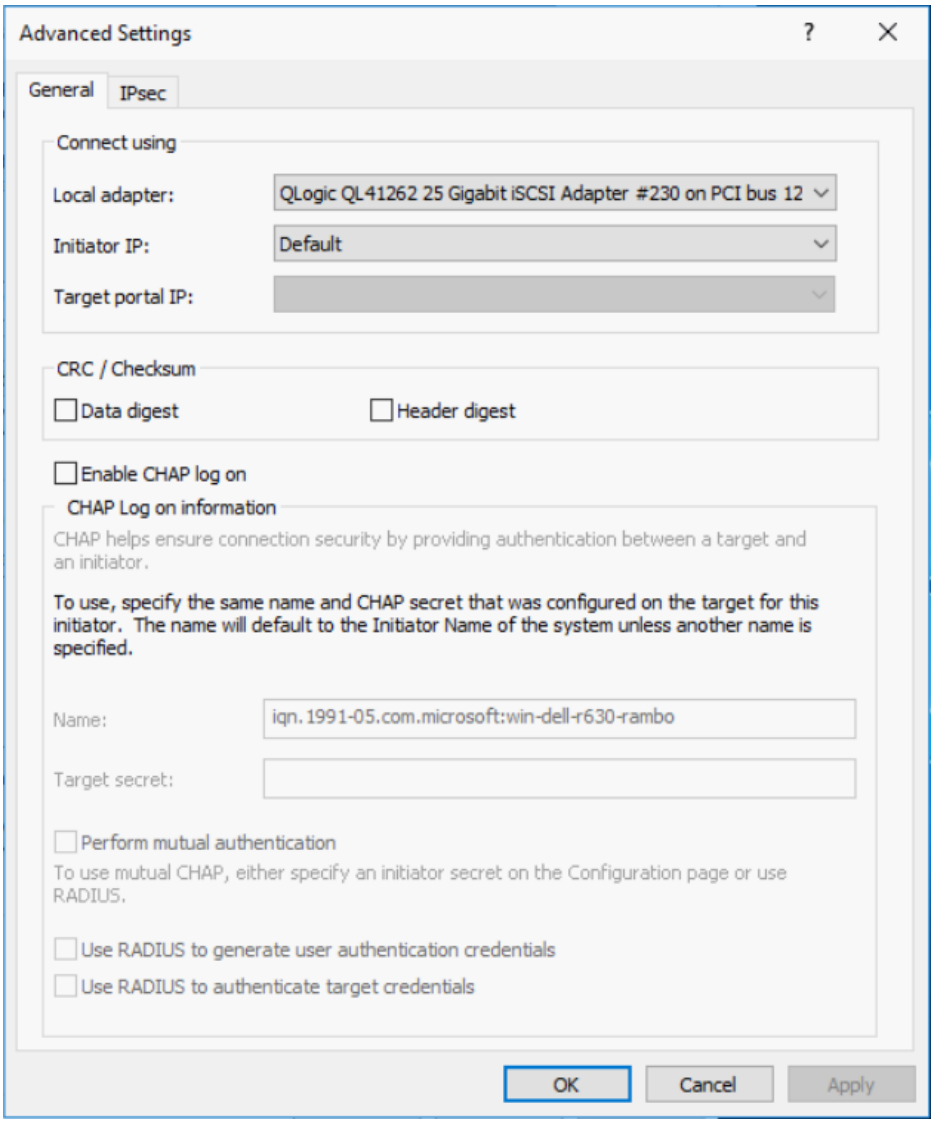


图 10-5. 选择启动器 IP 地址

7. 在 iSCSI Initiator Properties （iSCSI 启动器属性）、 Discovery （查找）页面上，单击 **OK** （确定）。

8. 单击 **Targets**（目标）选项卡，然后在 Targets（目标）页面（图 10-6）上，单击 **Connect**（连接）。

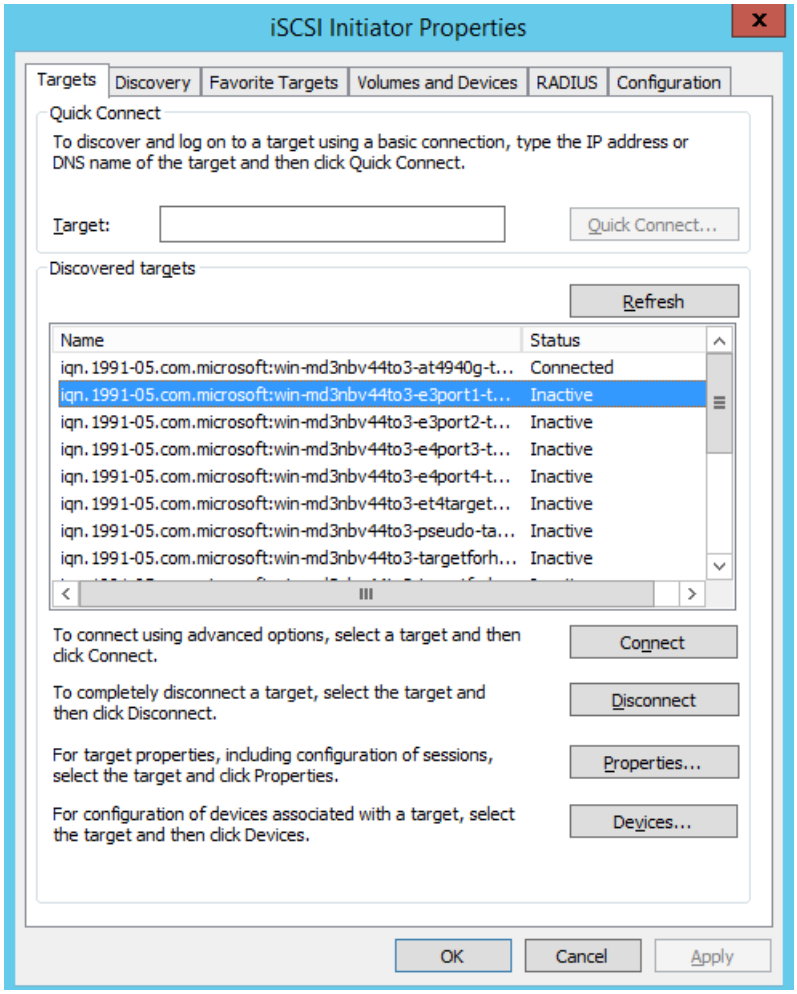


图 10-6. 连接到 iSCSI 目标

9. 在 Connect To Target （连接到目标）对话框 （图 10-7）上，单击 **Advanced** （高级）。

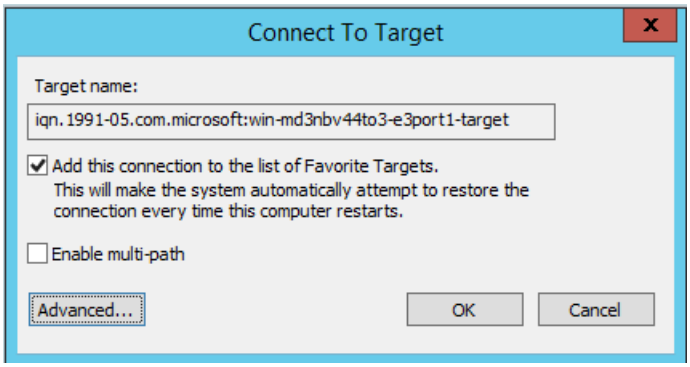


图 10-7. 连接到目标对话框

10. 在 Local Adapter （本地适配器）对话框中，选择 **QLogic <name or model> Adapter** （QLogic < 名称或型号 > 适配器），然后单击 **OK** （确定）。
11. 再次单击 **OK** （确定）关闭 Microsoft 启动器。
12. 要格式化 iSCSI 分区，请使用磁盘管理器。

注

组合功能的一些限制包括：

- 组合不支持 iSCSI 适配器。
- 组合不支持位于引导路径中的 NDIS 适配器。
- 组合支持不位于 iSCSI 引导路径中的 NDIS 适配器，但仅用于与交换机无关的 NIC 组类型。
- 与交换机有关的组合 （IEEE 802.3ad LACP 和通用 / 静态链路聚合（链路聚合）不能使用与交换机不相关的分区虚拟适配器。 IEEE 标准要求与交换机相关的组合 （IEEE 802.3ad LACP 和通用 / 静态链路聚合 （中继））模式按整个端口，而不是仅按 MAC 地址 （端口的一小部分）粒度工作。
- Microsoft 建议在 Windows Server 2012 及更高版本上使用其 in-OS NIC 组合服务，而不是任何适配器供应商专有的 NIC 组合驱动程序。

iSCSI 卸载常见问题

关于 iSCSI 卸载的一些常见问题包括：

问题： 如何为 iSCSI 卸载分配 IP 地址？

答案： 使用 QConvergeConsole GUI 中的 Configurations （配置）页面。

问题： 创建到目标的连接时应使用哪些工具？

答案： 使用 Microsoft iSCSI 软件启动器 （版本 2.08 或更高版本）。

问题： 怎样知道连接已卸载？

答案： 使用 Microsoft iSCSI 软件启动器。在命令行中，键入 `oiscsicli sessionlist`。从 **Initiator Name** （启动器名称），iSCSI 卸载的连接将显示以 `B06BDRV` 开始的条目。非卸载的连接将显示以 `Root` 开始的条目。

问题： 哪些配置应避免？

答案： IP 地址不能与 LAN 相同。

Windows Server 2012 R2、 2016 和 2019 iSCSI 引导安装

Windows Server 2012 R2、 Windows Server 2016 和 Windows Server 2019 支持在卸载或非卸载路径中引导和安装。Marvell 要求使用滑流 DVD，同时注入最新的 Marvell 驱动程序。请参阅 第 120 页上 “ 将适配器驱动程序注入 （滑流至） Windows 映像文件中 ”。

以下过程准备通过卸载或非卸载路径安装和引导的映像。

要设置 Windows Server 2012 R2/2016/2019 iSCSI 引导：

1. 从要引导的系统 （远程系统）上移除所有本地硬盘驱动器。
2. 通过按照第 120 页上 “ 将适配器驱动程序注入 （滑流至） Windows 映像文件中 ” 中的滑流步骤进行操作，准备 Windows OS 安装介质。
3. 将最新的 Marvell iSCSI 引导映像加载到适配器的 NVRAM 中。
4. 配置 iSCSI 目标以允许从远程设备连接。确保目标有足够磁盘空间安装新的 OS。
5. 配置 UEFI HII 以设置 iSCSI 引导类型 （卸载或非卸载）、正确的启动器和 iSCSI 引导的目标参数。
6. 保存设置并重新引导系统。远程系统应连接至 iSCSI 目标，然后从 DVD-ROM 设备引导。
7. 从 DVD 引导并开始安装。

8. 请按照屏幕说明进行操作。

   在显示可用于安装的磁盘列表的窗口中，应看到 iSCSI 目标磁盘。此目标是位于远程 iSCSI 目标中且通过 iSCSI 引导协议连接的磁盘。

9. 要继续 Windows Server 2012 R2/2016 安装，请单击 **Next**（下一步），然后按照屏幕说明进行操作。作为安装过程的一部分，服务器将进行多次重新引导。

10. 您应该在服务器引导至 OS 后，运行驱动程序安装程序以完成 Marvell 驱动程序和应用程序安装。

iSCSI 故障转储

41*xxx* 系列适配器 的非卸载和卸载 iSCSI 引导均支持故障转储功能。配置 iSCSI 故障转储生成无需任何额外的配置。

Linux 环境中的 iSCSI 卸载

Marvell FastLinQ 41*xxx* iSCSI 软件包含一个名为 `qedi.ko` (qedi) 的内核模块。qedi 模块依赖于 Linux 内核的其他部分来实现特定功能：

- **`qed.ko`** 是用于常见 Marvell FastLinQ 41*xxx* 硬件初始化例程的 Linux eCore 内核模块。
- **`scsi_transport_iscsi.ko`** 是用于上行调用和下行调用会话管理的 Linux iSCSI 传输库。
- **`libiscsi.ko`** 是协议数据单元 (PDU) 和任务处理，以及会话内存管理所需的 Linux iSCSI 库功能。
- **`iscsi_boot_sysfs.ko`** 是为导出 iSCSI 引导信息提供帮助的 Linux iSCSI sysfs 接口。
- **`uio.ko`** 是 Linux 用户空间 I/O 接口，用于 iscsiuio 的轻型 L2 内存映射。

必须先加载这些模块，然后 qedi 才能正常工作。否则，您可能会遇到“未解析的符号”错误。如果 qedi 模块安装在分发版更新路径中，则必需模块通过 modprobe 自动加载。

本节提供有关 Linux 中 iSCSI 卸载的以下信息：

- 与 bnx2i 的差异
- 配置 qedi.ko
- 在 Linux 中验证 iSCSI 接口

与 bnx2i 的差异

Marvell FastLinQ 41*xxx* 系列适配器 (iSCSI) 的驱动程序 qedi 与以前 Marvell 8400 系列适配器的 Marvell iSCSI 卸载驱动程序 bnx2i 之间存在一些重要的差异。其中一些差异包括：

- qedi 直接绑定至 CNA 公开的 PCI 功能。
- qedi 不会位于 net_device 的顶部。
- qedi 不依赖于网络驱动程序 （例如 bnx2x 和 cnic）。
- qedi 不依赖于 cnic，但依赖于 qed。
- qedi 使用 `iscsi_boot_sysfs.ko` 负责导出 sysfs 中的引导信息，而从 SAN 的 bnx2i 引导依赖于 `iscsi_ibft.ko` 模块导出引导信息。

配置 qedi.ko

qedi 驱动程序自动绑定至 CNA 公开的 iSCSI 功能，且通过 Open-iSCSI 工具进行目标发现和绑定。此功能和操作与 bnx2i 驱动程序类似。

要加载 `qedi.ko` 内核模块，请发出以下命令：

```
# modprobe qed
# modprobe libiscsi
# modprobe uio
# modprobe iscsi_boot_sysfs
# modprobe qedi
```

在 Linux 中验证 iSCSI 接口

安装和加载 qedi 内核模块后，必须验证是否正确检测到 iSCSI 接口。

要在 Linux 中验证 iSCSI 接口：

1. 要验证是否已主动加载 qedi 和关联的内核模块，请发出以下命令：

```
# lsmod | grep qedi
qedi                   114578   2
qed                    697989   1 qedi
uio                    19259    4 cnic,qedi
libiscsi               57233    2 qedi,bnx2i
scsi_transport_iscsi   99909    5 qedi,bnx2i,libiscsi
iscsi_boot_sysfs       16000    1 qedi
```

2. 要验证是否正确检测到 iSCSI 接口，请发出以下命令。在本示例中，检测到的两个 iSCSI CNA 设备具有 SCSI 主机编号 4 和 5。

```
# dmesg | grep qedi
[0000:00:00.0]:[qedi_init:3696]: QLogic iSCSI Offload Driver v8.15.6.0.
....
[0000:42:00.4]:[__qedi_probe:3563]:59: QLogic FastLinQ iSCSI Module qedi 8.15.6.0, FW 8.15.3.0
....
[0000:42:00.4]:[qedi_link_update:928]:59: Link Up event.
....
[0000:42:00.5]:[__qedi_probe:3563]:60: QLogic FastLinQ iSCSI Module qedi 8.15.6.0, FW 8.15.3.0
....
[0000:42:00.5]:[qedi_link_update:928]:59: Link Up event
```

3. 使用 Open-iSCSI 工具验证 IP 是否正确配置。发出以下命令：

```
# iscsiadm -m iface | grep qedi
qedi.00:0e:1e:c4:e1:6d
qedi,00:0e:1e:c4:e1:6d,192.168.101.227,<empty>,iqn.1994-05.com.redhat:534ca9b6adf
qedi.00:0e:1e:c4:e1:6c
qedi,00:0e:1e:c4:e1:6c,192.168.25.91,<empty>,iqn.1994-05.com.redhat:534ca9b6adf
```

4. 要确保 iscsiuio 服务正在运行，请发出以下命令：

```
# systemctl status iscsiuio.service
iscsiuio.service - iSCSI UserSpace I/O driver
Loaded: loaded (/usr/lib/systemd/system/iscsiuio.service; disabled; vendor preset: disabled)
Active: active (running) since Fri 2017-01-27 16:33:58 IST; 6 days ago
Docs: man:iscsiuio(8)
Process: 3745  ExecStart=/usr/sbin/iscsiuio (code=exited, status=0/SUCCESS)
    Main PID: 3747 (iscsiuio)
    CGroup: /system.slice/iscsiuio.service !--3747 /usr/sbin/iscsiuio
    Jan 27 16:33:58 localhost.localdomain systemd[1]: Starting iSCSI
UserSpace I/O driver...
    Jan 27 16:33:58 localhost.localdomain systemd[1]: Started iSCSI UserSpace I/O driver.
```

5. 要查找 iSCSI 目标，请发出 `iscsiadm` 命令：

```
#iscsiadm -m discovery -t st -p 192.168.25.100 -I qedi.00:0e:1e:c4:e1:6c
192.168.25.100:3260,1 iqn.2003-
04.com.sanblaze:virtualun.virtualun.target-05000007
192.168.25.100:3260,1 iqn.2003-04.com.sanblaze:virtualun.virtualun.target-05000012
192.168.25.100:3260,1 iqn.2003-04.com.sanblaze:virtualun.virtualun.target-0500000c
192.168.25.100:3260,1 iqn.2003-
04.com.sanblaze:virtualun.virtualun.target-05000001
192.168.25.100:3260,1 iqn.2003-04.com.sanblaze:virtualun.virtualun.target-05000002
```

6. 使用在 步骤 5 获取的 IQN 登录到 iSCSI 目标。要启动登录过程，请发出以下命令 （其中命令中最后一个字符为小写字母 “L”）：

```
#iscsiadm -m node -p 192.168.25.100 -T
iqn.2003-04.com.sanblaze:virtualun.virtualun.target-0000007 -l
Logging in to [iface: qedi.00:0e:1e:c4:e1:6c,
target:iqn.2003-04.com.sanblaze:virtualun.virtualun.target-05000007, portal:192.168.25.100,3260]
(multiple)
Login to [iface: qedi.00:0e:1e:c4:e1:6c, target:iqn.2003-
04.com.sanblaze:virtualun.virtualun.target-05000007, portal:192.168.25.100,3260] successful.
```

7. 要验证 iSCSI 会话已创建，请发出以下命令：

```
# iscsiadm -m session
qedi: [297] 192.168.25.100:3260,1
iqn.2003-04.com.sanblaze:virtualun.virtualun.target-05000007 (non-flash)
```

8. 要检查 iSCSI 设备，请发出 `iscsiadm` 命令：

```
# iscsiadm -m session -P3
...
************************
Attached SCSI devices:
************************
Host Number: 59 State: running
scsi59 Channel 00 Id 0 Lun: 0
Attached scsi disk sdb   State: running scsi59 Channel 00 Id 0 Lun: 1
Attached scsi disk sdc   State: running scsi59 Channel 00 Id 0 Lun: 2
Attached scsi disk sdd   State: running scsi59 Channel 00 Id 0 Lun: 3
Attached scsi disk sde   State: running scsi59 Channel 00 Id 0 Lun: 4
Attached scsi disk sdf   State: running
```

有关高级目标配置，请参阅 Open-iSCSI 自述文件，网址为：

https://github.com/open-iscsi/open-iscsi/blob/master/README

11 FCoE 配置

本章提供以下以太网光纤信道 (FCoE) 配置信息：

- 第 210 页上 “ 配置 Linux FCoE 卸载 ”

注

所有 41*xxx* 系列适配器都支持 FCoE 卸载。某些 FCoE 功能在当前版本中可能并未完全启用。有关详细信息，请参阅 附录 D 功能约束。

要启用 iSCSI 卸载模式，请参阅位于 https://www.marvell.com/documents/5aa5otcbkr0im3ynera3/ 的 *Application Note, Enabling Storage Offloads on Dell and Marvell FastLinQ 41000 Series Adapters*。

有关从 SAN 的 FCoE 引导信息，请参阅第 6 章 从 SAN 引导配置。

配置 Linux FCoE 卸载

Marvell FastLinQ 41*xxx* 系列适配器 FCoE 软件包含一个名为 qedf.ko (qedf) 的内核模块。 qedf 模块依赖于 Linux 内核的其他部分来实现特定功能：

- **`qed.ko`** 是用于常见 Marvell FastLinQ 41*xxx* 硬件初始化例程的 Linux eCore 内核模块。
- **`libfcoe.ko`** 是执行 FCoE 转发器 (FCF) 请求和 FCoE 初始化协议 (FIP) 结构登录 (FLOGI) 所需的 Linux FCoE 内核库。
- **`libfc.ko`** 是多项功能所需的 Linux FC 内核库，这些功能包括：
  - 名称服务器登录和注册
  - rport 会话管理
- **`scsi_transport_fc.ko`** 是用于远程端口和 SCSI 目标管理的 Linux FC SCSI 传输库。

必须加载这些模块后 qedf 才能正常工作，否则可导致 “ 未解析的符号 ” 等错误。如果 qedf 模块安装在分发版更新路径中，则必需模块通过 modprobe 自动加载。Marvell FastLinQ 41*xxx* 系列适配器支持 FCoE 卸载。

本节提供有关 Linux 中 FCoE 卸载的以下信息：

qedf 与 bnx2fc 之间的差异

Marvell FastLinQ 41*xxx* 10/25GbE 控制器 (FCoE) 的驱动程序 qedf 与以前的 Marvell FCoE 卸载驱动程序 bnx2fc 之间存在显著的差异。差异包括：

- qedf 直接绑定至 CNA 公开的 PCI 功能。
- qedf 无需 open-fcoe 用户空间工具（fipvlan、fcoemon、fcoeadm）即可启动查找。
- qedf 直接发出 FIP vLAN 请求而无需 fipvlan 公用程序。
- qedf 无需 fipvlan 为 fcoemon 创建的 FCoE 接口。
- qedf 不会位于 net_device 的顶部。
- qedf 不依赖于网络驱动程序 （例如 bnx2x 和 cnic）。
- qedf 将在链路正常工作时自动启动 FCoE 查找 （因为它不依赖于 fipvlan 或 fcoemon 进行 FCoE 接口创建）。

注

FCoE 接口不再位于网络接口之上。 qedf 驱动程序自动创建独立于网络接口的 FCoE 接口。因此， FCoE 接口不会显示在安装程序的 FCoE 接口对话框中。相反，磁盘会自动显示为 SCSI 磁盘，与光纤信道驱动程序的工作方式类似。

配置 qedf.ko

qedf.ko 无需显式配置。驱动程序会自动绑定至 CNA 公开的 FCoE 功能并开始查找。相比于较旧的 bnx2fc 驱动程序，此功能与 Marvell FC 驱动程序 qla2xx 的功能和运行类似。

注

有关 FastLinQ 驱动程序安装的更多信息，请参阅第 3 章 驱动程序安装。

加载 qedf.ko 内核模块会执行以下命令：

```
# modprobe qed
# modprobe libfcoe
# modprobe qedf
```

在 Linux 中验证 FCoE 设备

按照以下步骤操作，验证在安装和加载 qedf 内核模块后是否正确检测到 FCoE 设备。

要在 Linux 中验证 FCoE 设备：

1. 检查 lsmod，验证是否已加载 qedf 和关联的内核模块：

```
# lsmod | grep qedf
69632   1   qedf libfc
143360  2   qedf,libfcoe scsi_transport_fc
65536   2   qedf,libfc qed
806912  1   qedf scsi_mod
262144  14  sg,hpsa,qedf,scsi_dh_alua,scsi_dh_rdac,dm_multipath,
scsi_transport_fc,scsi_transport_sas,libfc,scsi_transport_iscsi,scsi_dh_emc,
libata,sd_mod,sr_mod
```

2. 检查 dmesg，验证是否正确检测到 FCoE 设备。在本示例中，检测到的两个 FCoE CNA 设备具有 SCSI 主机编号 4 和 5。

```
# dmesg | grep qedf
[  235.321185] [0000:00:00.0]: [qedf_init:3728]: QLogic FCoE Offload Driver
v8.18.8.0.
....
[  235.322253] [0000:21:00.2]: [__qedf_probe:3142]:4: QLogic FastLinQ FCoE
Module qedf 8.18.8.0, FW 8.18.10.0
[  235.606443] scsi host4: qedf
....
[  235.624337] [0000:21:00.3]: [__qedf_probe:3142]:5: QLogic FastLinQ FCoE
Module qedf 8.18.8.0, FW 8.18.10.0
[  235.886681] scsi host5: qedf
....
[  243.991851] [0000:21:00.3]: [qedf_link_update:489]:5: LINK UP (40 GB/s).
```

3. 使用 `lsscsi` 或 `lsblk -S` 命令检查找到的 FCoE 设备。每个命令的示例如下所示。

```
# lsscsi
[0:2:0:0]    disk    DELL     PERC H700        2.10  /dev/sda
[2:0:0:0]    cd/dvd  TEAC     DVD-ROM DV-28SW  R.2A  /dev/sr0
[151:0:0:0]  disk    HP       P2000G3 FC/iSCSI T252  /dev/sdb
[151:0:0:1]  disk    HP       P2000G3 FC/iSCSI T252  /dev/sdc
[151:0:0:2]  disk    HP       P2000G3 FC/iSCSI T252  /dev/sdd
[151:0:0:3]  disk    HP       P2000G3 FC/iSCSI T252  /dev/sde
[151:0:0:4]  disk    HP       P2000G3 FC/iSCSI T252  /dev/sdf
```

```
# lsblk -S
NAME HCTL       TYPE  VENDOR    MODEL             REV TRAN
sdb  5:0:0:0    disk  SANBlaze  VLUN P2T1L0       V7.3 fc
sdc  5:0:0:1    disk  SANBlaze  VLUN P2T1L1       V7.3 fc
sdd  5:0:0:2    disk  SANBlaze  VLUN P2T1L2       V7.3 fc
sde  5:0:0:3    disk  SANBlaze  VLUN P2T1L3       V7.3 fc
sdf  5:0:0:4    disk  SANBlaze  VLUN P2T1L4       V7.3 fc
sdg  5:0:0:5    disk  SANBlaze  VLUN P2T1L5       V7.3 fc
sdh  5:0:0:6    disk  SANBlaze  VLUN P2T1L6       V7.3 fc
sdi  5:0:0:7    disk  SANBlaze  VLUN P2T1L7       V7.3 fc
sdj  5:0:0:8    disk  SANBlaze  VLUN P2T1L8       V7.3 fc
sdk  5:0:0:9    disk  SANBlaze  VLUN P2T1L9       V7.3 fc
```

主机的配置信息位于 `/sys/class/fc_host/hostX`，其中 `X` 是 SCSI 主机的编号。在上面的示例中，`X` 是 4。`hostX` 文件包含 FCoE 功能的属性，例如全局端口名称和结构 ID。

12 SR-IOV 配置

单根输入 / 输出虚拟化 (SR-IOV) 是一种 PCI SIG 规格，使单个 PCI Express (PCIe) 设备能够显示为多个单独的物理 PCIe 设备。 SR-IOV 允许隔离 PCIe 资源以实现性能、互操作性和可管理性。

注

某些 SR-IOV 功能在当前版本中可能并未完全启用。

本章提供以下内容的说明：

- 在 Windows 上配置 SR-IOV
- 第 221 页上 “ 在 Linux 上配置 SR-IOV”
- 第 228 页上 “ 在 VMware 上配置 SR-IOV”

在 Windows 上配置 SR-IOV

要在 Windows 上配置 SR-IOV：

1. 访问服务器 BIOS System Setup （BIOS 系统设置），然后单击 **System BIOS Settings** （系统 BIOS 设置）。
2. 在 System BIOS Settings （系统 BIOS 设置）页面上，单击 **Integrated Devices** （集成式设备）。
3. 在 Integrated Devices （集成式设备）页面 （图 12-1）上：
   a. 将 **SR-IOV Global Enable** （SR-IOV 全局启用）选项设置为 **Enabled** （已启用）。
   b. 单击 **Back** （后退）。

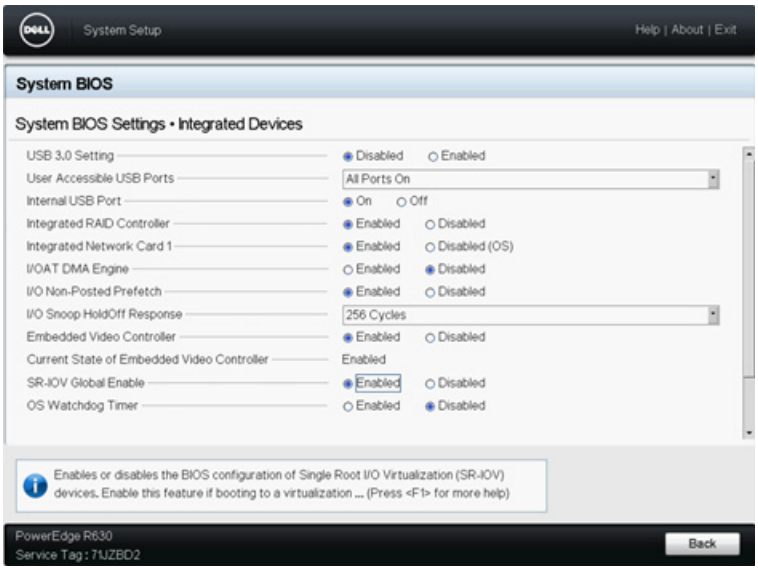


图 12-1. SR-IOV 的系统设置：集成式设备

4. 在所选适配器的 Main Configuration Page （主要配置页面）上，单击 **Device Level Configuration** （设备级配置）。
5. 在 Main Configuration Page - Device Level Configuration （主要配置页面 - 设备级配置）（图 12-2）上：
   a. 如果您使用的是 NPAR 模式，请将 **Virtualization Mode** （虚拟化模式）设置为 **SR-IOV** 或 **NPAR+SR-IOV**。
   b. 单击 **Back** （后退）。

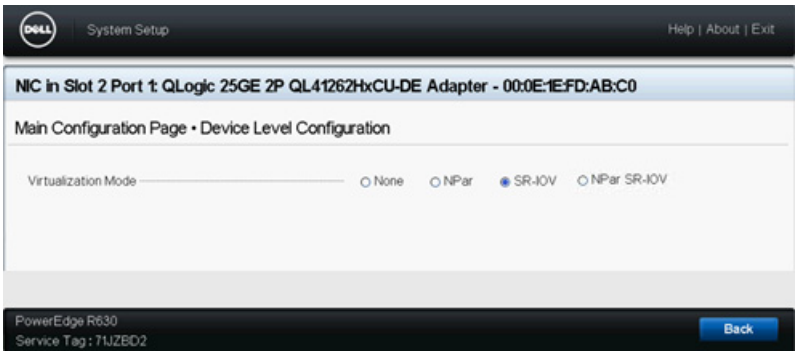


图 12-2. SR-IOV 的系统设置：设备级配置

6. 在 Main Configuration Page （主要配置页面）上，单击 **Finish** （完成）。
7. 在 Warning - Saving Changes （警告 - 保存更改）消息框中，单击 **Yes** （是）保存配置。

8. 在 Success - Saving Changes （成功 - 保存更改）消息框中，单击 **OK** （确定）。
9. 要在微型端口适配器上启用 SR-IOV：
   a. 访问设备管理器。
   b. 打开微型端口适配器属性，然后单击 **Advanced** （高级）选项卡。
   c. 在 Advanced （高级）属性页面 （图 12-3）的 **Property** （属性）下，选择 **SR-IOV**，然后将值设置为 **Enabled** （已启用）。
   d. 单击 **OK** （确定）。

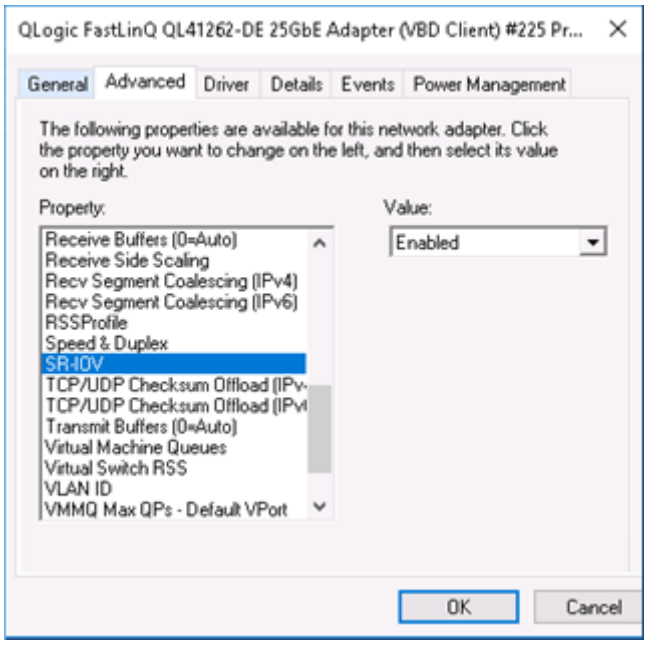


图 12-3. 适配器属性，高级：启用 SR-IOV

10. 要使用 SR-IOV 创建虚拟机交换机 (vSwitch)(第 217 页上 图 12-4）：
    a. 启动 Hyper-V 管理器。
    b. 选择 **Virtual Switch Manager** （虚拟交换机管理器）。
    c. 在 **Name** （名称）框中，键入虚拟交换机的名称。
    d. 在 **Connection type** （连接类型）下，选择 **External network** （外部网络）。
    e. 选择 **Enable single-root I/O virtualization (SR-IOV)** （启用单根 I/O 虚拟化 (SR-IOV)）复选框，然后单击 **Apply** （应用）。

注

请确保在创建 vSwitch 时启用 SR-IOV。创建 vSwitch 后此选项不可用。

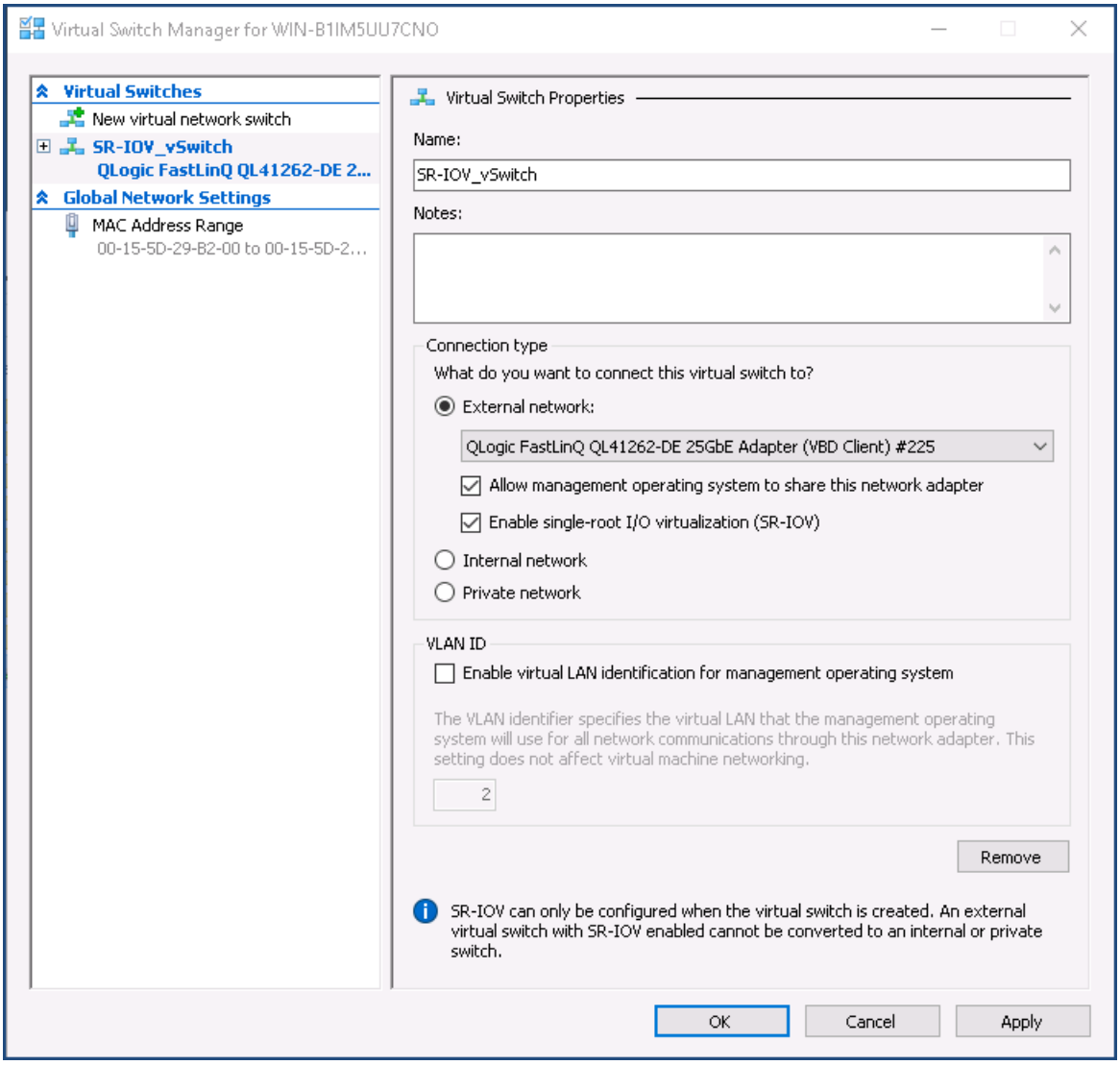


图 12-4. 虚拟交换机管理器：启用 SR-IOV

f. Apply Networking Changes （应用网络更改）消息框告知您 **Pending changes may disrupt network connectivity** （待处理更改可能会中断网络连接）。要保存您的更改并继续，请单击 **Yes** （是）。

11. 要获取虚拟机交换机功能，请发出以下 Windows PowerShell 命令：

```
PS C:\Users\Administrator> Get-VMSwitch -Name SR-IOV_vSwitch | fl
```

`Get-VMSwitch` 命令的输出将包括以下 SR-IOV 功能：

```
IovVirtualFunctionCount              : 80
IovVirtualFunctionsInUse             : 1
```

12. 要创建虚拟机 (VM) 并导出 VM 中的虚拟功能 (VF)：
    a. 创建一个虚拟机。
    b. 将 VMNetworkadapter 添加到虚拟机。
    c. 将虚拟交换机分配给 VMNetworkadapter。
    d. 在 Settings for VM <VM_Name> （VM <VM_ 名称 > 的设置）对话框（图 12-5）中， Hardware Acceleration （硬件加速）页面的 **Single-root I/O virtualization** （单根 I/O 虚拟化）下，选择 **Enable SR-IOV** （启用 SR-IOV）复选框，然后单击 **OK** （确定）。

注

创建虚拟适配器连接后，可在任何时候 （即使在流量运行的时候）启用或禁用 SR-IOV 设置。

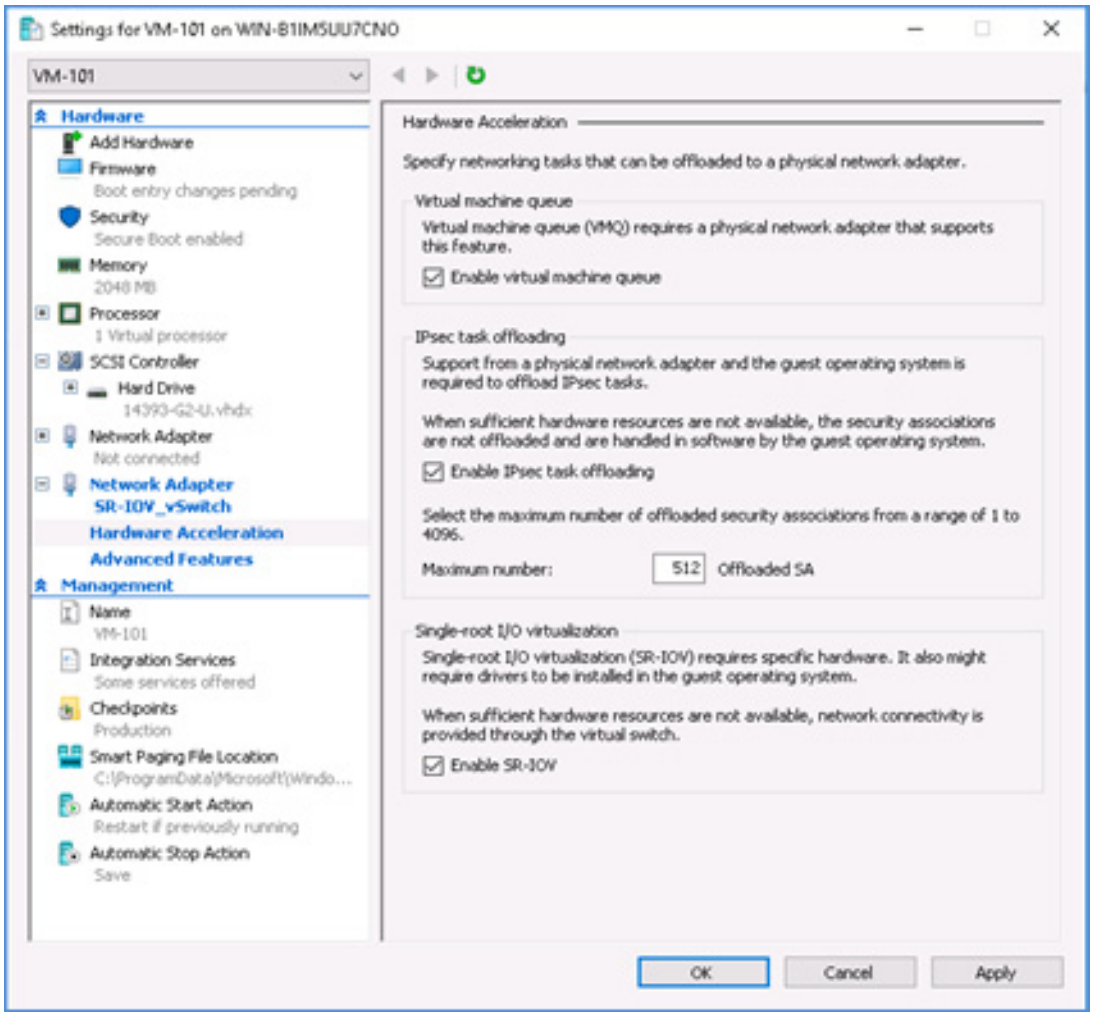


图 12-5. VM 的设置：启用 SR-IOV

13. 为 VM 中检测到的适配器安装 Marvell 驱动程序。为主机 OS 使用可从供应商处获得的最新驱动程序 （请勿使用内建驱动程序）。

注

请确保在 VM 和主机系统中使用相同的驱动程序包。例如，在 Windows VM 和 Windows Hyper-V 主机使用相同 qeVBD 和 qeND 驱动程序版本。

安装驱动程序后，适配器在 VM 中列出。图 12-6 显示一个示例。

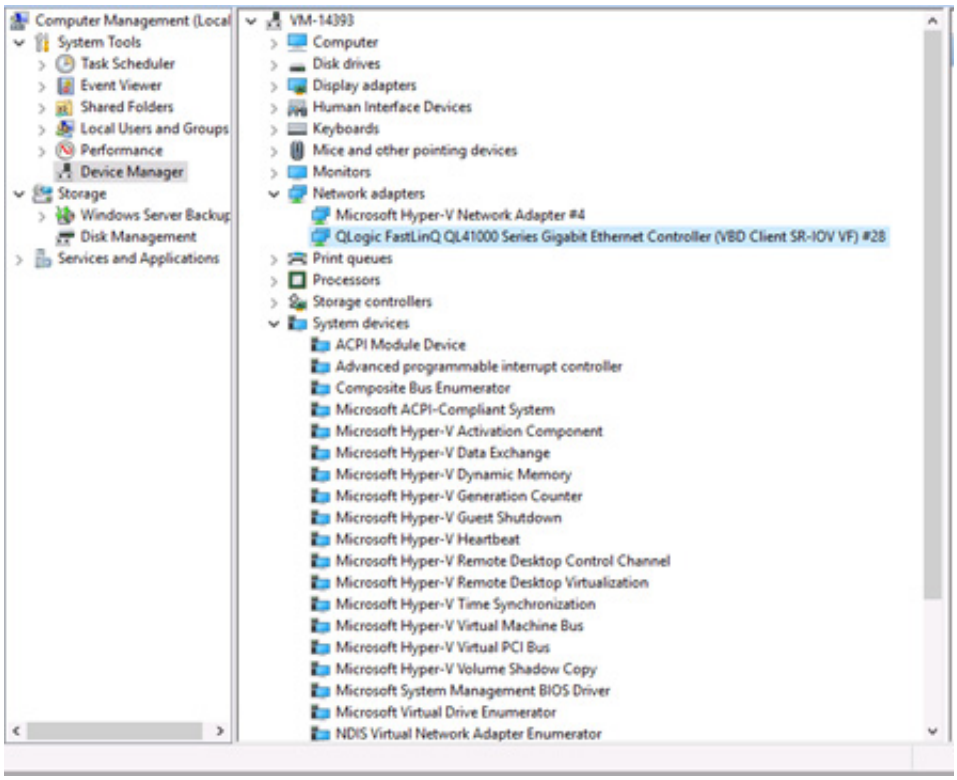


图 12-6. 设备管理器：带 QLogic 适配器的 VM

14. 要查看 SR-IOV VF 详细信息，请发出以下 Windows PowerShell 命令：

```
PS C:\Users\Administrator> Get-NetadapterSriovVf
```

图 12-7 显示示例输出。

```
PS C:\Users\Administrator>
PS C:\Users\Administrator> Get-NetAdapterSriovVf

Name                      FunctionID VPortID MacAddress        VmID                                 VmFriendlyName
----                      ---------- ------- ----------        ----                                 --------------
Ethernet 10               0          {2}     00-15-5D-29-B2-01 51F01C52-CDC6-4932-A95E-86D... VM-101

PS C:\Users\Administrator> _
```

图 12-7. Windows PowerShell 命令：Get-NetadapterSriovVf

在 Linux 上配置 SR-IOV

要在 Linux 上配置 SR-IOV：

1. 访问服务器 BIOS System Setup （BIOS 系统设置），然后单击 **System BIOS Settings** （系统 BIOS 设置）。
2. 在 System BIOS Settings （系统 BIOS 设置）页面上，单击 **Integrated Devices** （集成式设备）。
3. 在 System Integrated Devices （系统集成式设备）页面 （请参阅第 215 页上 图 12-1）上：
   a. 将 **SR-IOV Global Enable** （SR-IOV 全局启用）选项设置为 **Enabled** （已启用）。
   b. 单击 **Back** （后退）。
4. 在 System BIOS Settings （系统 BIOS 设置）页面上，单击 **Processor Settings** （处理器设置）。
5. 在 Processor Settings （处理器设置）（图 12-8）页面上：
   a. 将 **Virtualization Technology** （虚拟化技术）选项设置为 **Enabled** （已启用）。
   b. 单击 **Back** （后退）。

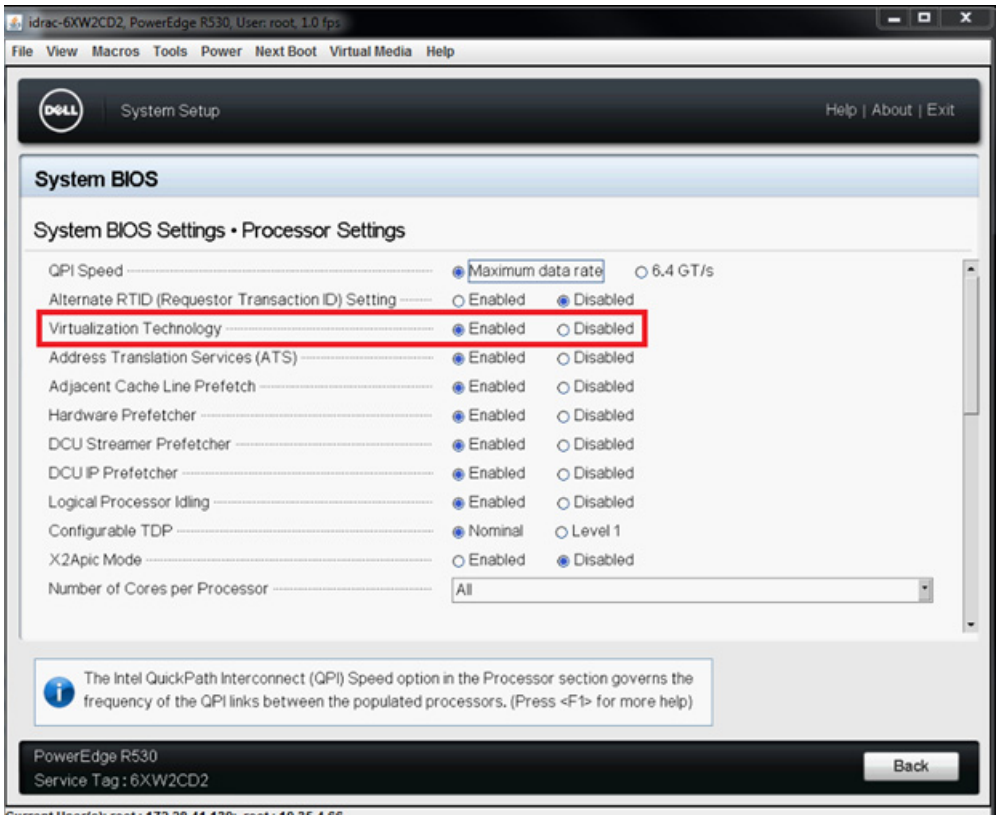


图 12-8. 系统设置：SR-IOV 的处理器设置

6. 在 System Setup （系统设置）页面上，选择 **Device Settings** （设备设置）。
7. 在 Device Settings （设备设置）页面中，为 Marvell 适配器选择 **Port 1** （端口 1）。
8. 在 Device Level Configuration （设备级配置）页面 （图 12-9）上：
   a. 将 **Virtualization Mode** （虚拟化模式）设置为 **SR-IOV**。
   b. 单击 **Back** （后退）。

图 12-9. SR-IOV 的系统设置：集成式设备

9. 在 Main Configuration Page （主要配置页面）上，单击 **Finish** （完成），保存设置，然后重新引导系统。
10. 要启用并验证虚拟化：
    a. 打开 `grub.conf` 文件并配置 `iommu` 参数，如图 12-10 中所示。（有关详细信息，请参阅 第 227 页上 “ 启用以基于 UEFI 的 Linux OS 安装中 SR-IOV 的 IOMMU”。）
       - 对于基于英特尔的系统，请添加 `intel_iommu=on`。
       - 对于基于 AMD 的系统，请添加 `amd_iommu=on`。

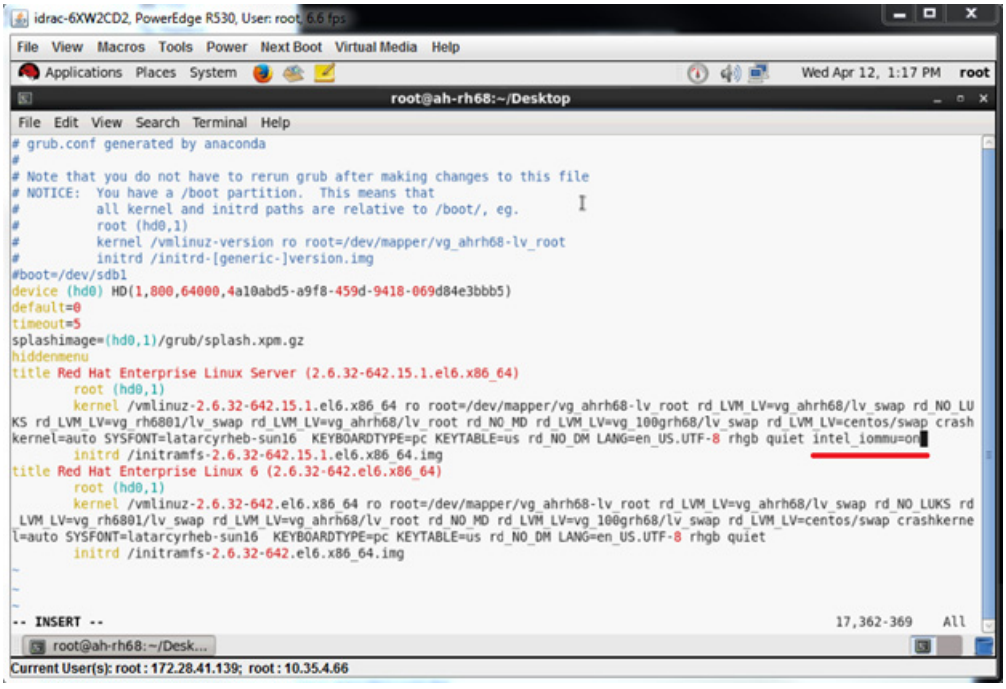


图 12-10. 为 SR-IOV 编辑 grub.conf 文件

    b. 保存 `grub.conf` 文件，然后重新引导系统。
    c. 要验证更改是否已生效，请发出以下命令：

```
dmesg | grep  -i iommu
```

       应显示成功的输入 - 输出内存管理单元 (IOMMU) 命令输出，例如；

```
Intel-IOMMU: enabled
```

    d. 要查看 VF 详细信息 （VF 数和 VF 总数），请发出以下命令：

```
find /sys/|grep -i sriov
```

11. 对于特定端口，将启用 VF 数量。

    a. 请发出以下命令启用，例如在 PCI 实例 04:00.0 （总线 4，设备 0，功能 0）上启用 8 个 VF：

    ```
    [root@ah-rh68 ~]# echo 8 >
    /sys/devices/pci0000:00/0000:00:02.0/0000:04:00.0/
    sriov_numvfs
    ```

    b. 查看命令输出 (图 12-11) 以确认通过 7 在总线 4，设备 2（0000:00:02.0 参数），功能 0 中已创建有实际 VF。请注意 PF （在此示例中 ID 为 8070) 上的实际设备 ID 与 VF （在此示例中 ID 为 8090）上的不同。

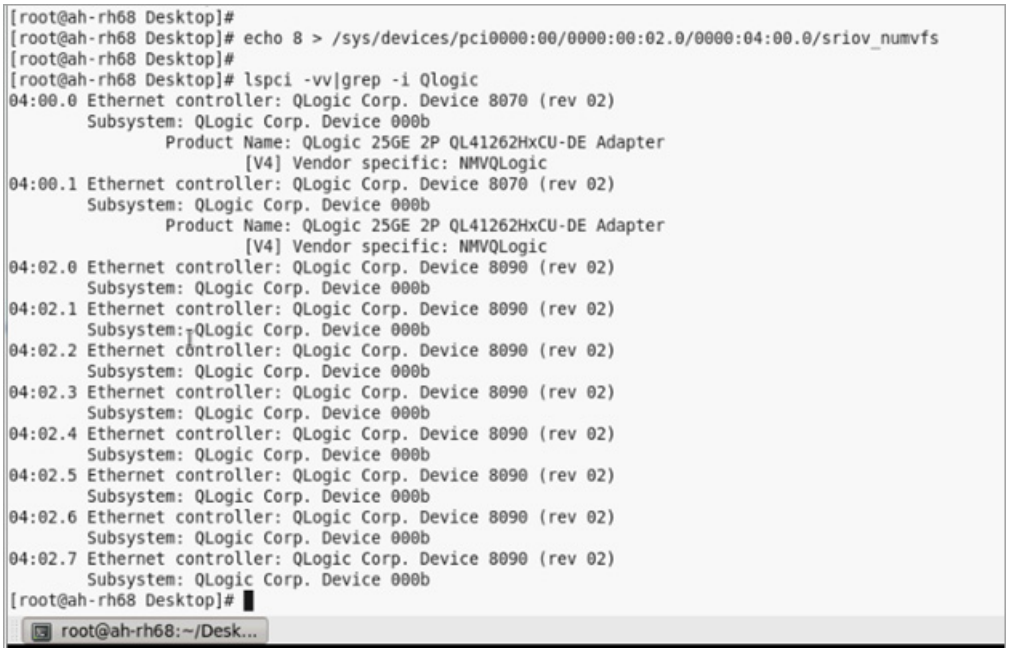


图 12-11. sriov_numvfs 的命令输出

12. 要查看所有 PF 和 VF 接口的列表，请发出以下命令：

    ```
    # ip link show | grep -i vf -b2
    ```

图 12-12 显示示例输出。

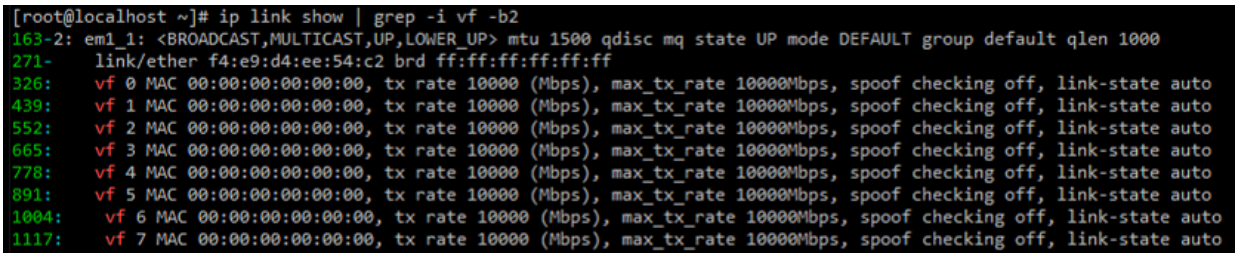


图 12-12. ip link show 命令的命令输出

13. 分配并验证 MAC 地址：
    a. 要将 MAC 地址分配给 VF，请发出以下命令：

       `ip link set <pf device> vf <vf index> mac <mac address>`

    b. 确保 VF 接口正常工作并以分配的 MAC 地址运行。

14. 关闭 VM 的电源并连接 VF。（某些 OS 支持热插拔 VF 到 VM。）

    a. 在 Virtual Machine （虚拟机）对话框 （图 12-13）中，单击 **Add Hardware** （添加硬件）。

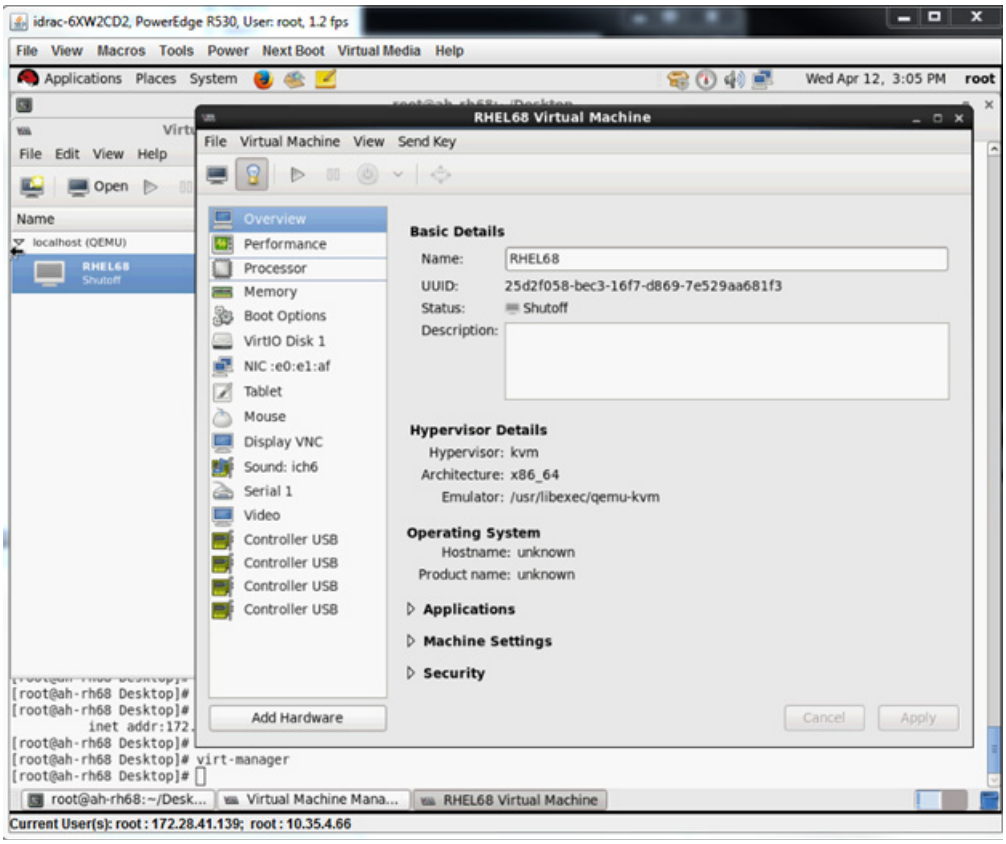


图 12-13. RHEL68 虚拟机

    b. 在 Add New Virtual Hardware（添加新虚拟硬件）对话框 （图 12-14）的左侧窗格中，单击 **PCI Host Device** （PCI 主机设备）。

    c. 在右侧窗格中，选择一个主机设备。

    d. 单击 **Finish** （完成）。

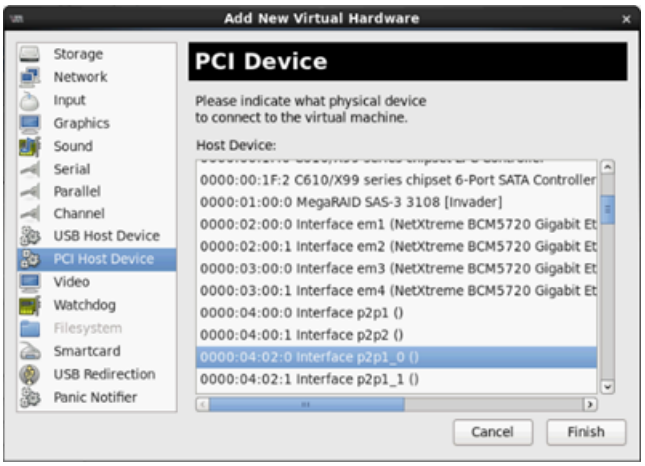


图 12-14. 添加新虚拟硬件

15. 打开 VM 的电源，然后发出以下命令：

    `check lspci -vv|grep -I ether`

16. 为 VM 中检测到的适配器安装驱动程序。为主机 OS 使用可从供应商处获得的最新驱动程序 （请勿使用内建驱动程序）。主机和 VM 上必须安装相同的驱动程序版本。

17. 根据需要，在 VM 中添加更多 VF。

启用以基于 UEFI 的 Linux OS 安装中 SR-IOV 的 IOMMU

请对 Linux OS 执行正确的步骤。

注

对于 AMD 系统，替换 `intel_iommu=on` 为 `amd_iommu=on`。

要在 RHEL 6.*x* 上启用 SR-IOV 的 IOMMU：

- 在 `/boot/efi/EFI/redhat/grub.conf` 文件中，定位内核行，并附加 `intel_iommu=on` 引导参数。

要在 RHEL 7.*x* 及其更高版本上启用 SR-IOV 的 IOMMU：

1. 在 `/etc/default/grub` 文件中，定位 `GRUB_CMDLINE_LINUX`，然后附加 `intel_iommu=on` 引导参数。

2. 要升级 grub 配置文件，请发出以下命令：

   `grub2-mkconfig -o /boot/efi/EFI/redhat/grub.cfg`

要在 SLES 12.*x* 上启用 SR-IOV 的 IOMMU：

1. 在 `/etc/default/grub` 文件中，定位 `GRUB_CMDLINE_LINUX_DEFAULT`，然后附加 `intel_iommu=on` 引导参数。
2. 要升级 grub 配置文件，请发出以下命令：

   `grub2-mkconfig -o /boot/grub2/grub.cfg`

要在 SLES 15.*x* 及其更高版本上启用 SR-IOV 的 IOMMU：

1. 在 `/etc/default/grub` 文件中，定位 `GRUB_CMDLINE_LINUX_DEFAULT`，然后附加 `intel_iommu=on` 引导参数。
2. 要升级 grub 配置文件，请发出以下命令：

   `grub2-mkconfig -o /boot/efi/EFI/sles/grub.cfg`

在 VMware 上配置 SR-IOV

要在 VMware 上配置 SR-IOV：

1. 访问服务器 BIOS System Setup （BIOS 系统设置），然后单击 **System BIOS Settings** （系统 BIOS 设置）。
2. 在 System BIOS Settings （系统 BIOS 设置）页面上，单击 **Integrated Devices** （集成式设备）。
3. 在 Integrated Devices （集成式设备）页面 （请参阅第 215 页上 图 12-1）上：
   a. 将 **SR-IOV Global Enable** （SR-IOV 全局启用）选项设置为 **Enabled** （已启用）。
   b. 单击 **Back** （后退）。
4. 在 System Setup （系统设置）窗口中，单击 **Device Settings** （设备设置）。
5. 在 Device Settings （设备设置）页面上，选择用于 25G 41*xxx* 系列适配器的端口。
6. 在 Device Level Configuration （设备级配置）页面 （请参阅第 215 页上 图 12-2）上：
   a. 将 **Virtualization Mode** （虚拟化模式）设置为 **SR-IOV**。
   b. 单击 **Back** （后退）。
7. 在 Main Configuration Page （主要配置页面）上，单击 **Finish** （完成）。

8. 保存配置设置并重新引导系统。
9. 要启用每一端口 VF 所需数量 （在此示例中，双端口适配器的每一端口数量为 16），请发出以下命令：

   "esxcfg-module -s "max_vfs=16,16" qedentv"

   注

   41*xxx* 系列适配器 的每一以太网功能必须有其各自条目。

10. 重新引导主机。
11. 要验证模块级别的更改是否已完成，请发出以下命令：

    "esxcfg-module -g qedentv"

    ```
    [root@localhost:~] esxcfg-module -g qedentv
    qedentv enabled = 1 options = 'max_vfs=16,16'
    ```

12. 要验证是否创建了实际 VF，请发出 `lspci` 命令如下：

```
[root@localhost:~] lspci | grep -i QLogic | grep -i 'ethernet\|network' | more
0000:05:00.0 Network controller: QLogic Corp. QLogic FastLinQ QL41xxx 10/25
GbE Ethernet Adapter [vmnic6]
0000:05:00.1 Network controller: QLogic Corp. QLogic FastLinQ QL41xxx 10/25
GbE Ethernet Adapter [vmnic7]
0000:05:02.0 Network controller: QLogic Corp. QLogic FastLinQ QL41xxx Series
10/25 GbE Controller (SR-IOV VF) [PF_0.5.0_VF_0]
0000:05:02.1 Network controller: QLogic Corp. QLogic FastLinQ QL41xxx Series
10/25 GbE Controller (SR-IOV VF) [PF_0.5.0_VF_1]
0000:05:02.2 Network controller: QLogic Corp. QLogic FastLinQ QL41xxx Series
10/25 GbE Controller (SR-IOV VF) [PF_0.5.0_VF_2]
0000:05:02.3 Network controller: QLogic Corp. QLogic FastLinQ QL41xxx Series
10/25 GbE Controller (SR-IOV VF) [PF_0.5.0_VF_3]
.
.
.
0000:05:03.7 Network controller: QLogic Corp. QLogic FastLinQ QL41xxx Series
10/25 GbE Controller (SR-IOV VF) [PF_0.5.0_VF_15]
0000:05:0e.0 Network controller: QLogic Corp. QLogic FastLinQ QL41xxx Series
10/25 GbE Controller (SR-IOV VF) [PF_0.5.1_VF_0]
0000:05:0e.1 Network controller: QLogic Corp. QLogic FastLinQ QL41xxx Series
10/25 GbE Controller (SR-IOV VF) [PF_0.5.1_VF_1]
0000:05:0e.2 Network controller: QLogic Corp. QLogic FastLinQ QL41xxx Series
10/25 GbE Controller (SR-IOV VF) [PF_0.5.1_VF_2]
```

```
0000:05:0e.3 Network controller: QLogic Corp. QLogic FastLinQ QL41xxx Series
10/25 GbE Controller (SR-IOV VF) [PF_0.5.1_VF_3]
.
.
.
0000:05:0f.6 Network controller: QLogic Corp. QLogic FastLinQ QL41xxx Series
10/25 GbE Controller (SR-IOV VF) [PF_0.5.1_VF_14]
0000:05:0f.7 Network controller: QLogic Corp. QLogic FastLinQ QL41xxx Series
10/25 GbE Controller (SR-IOV VF) [PF_0.5.1_VF_15]
```

13. 请执行以下操作将 VF 连接到 VM：
    a. 关闭 VM 的电源并连接 VF。（某些 OS 支持热插拔 VF 到 VM。）
    b. 将主机添加到 VMware vCenter Server 虚拟设备 (vCSA)。
    c. 单击 VM 的 **Edit Settings** （编辑设置）。
14. 请如下填写 Edit Settings （编辑设置）对话框 （图 12-15）：
    a. 在 **New Device** （新设备）框中，选择 **Network** （网络），然后单击 **Add** （添加）。
    b. 对于 **Adapter Type** （适配器类型），选择 **SR-IOV Passthrough** （SR-IOV 直通）。
    c. 对于 **Physical Function** （物理功能），选择 Marvell VF。
    d. 要保存您的配置更改并关闭此对话框，请单击 **OK** （确定）。

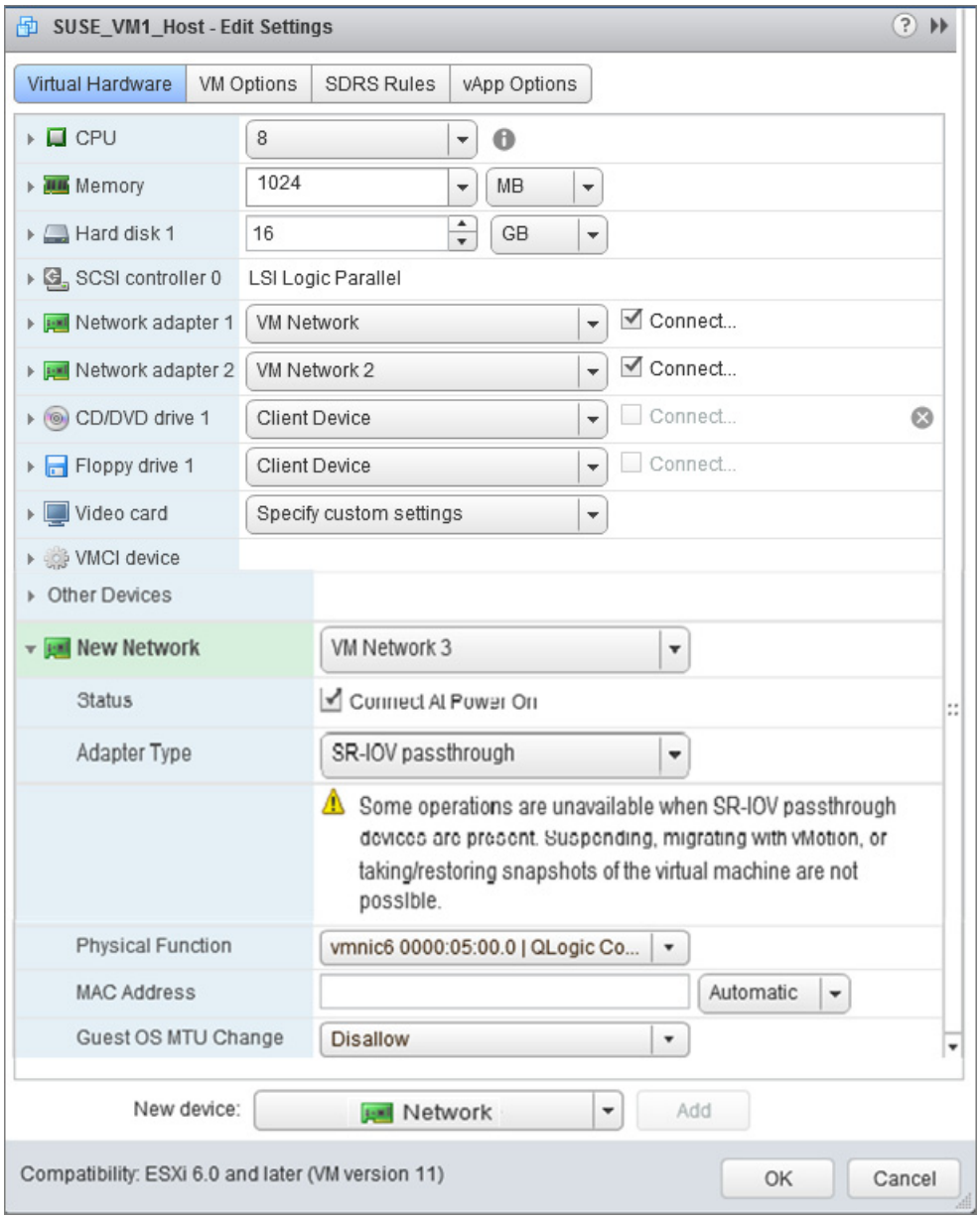


图 12-15. VMware 主机编辑设置

15. 要验证每个端口的 VF，请如下发出 `esxcli` 命令：

```
[root@localhost:~] esxcli network sriovnic vf list -n vmnic6
VF ID  Active  PCI Address  Owner World ID
-----  ------  -----------  --------------
    0    true  005:02.0     60591
    1    true  005:02.1     60591
```

```
 2   false  005:02.2      -
 3   false  005:02.3      -
 4   false  005:02.4      -
 5   false  005:02.5      -
 6   false  005:02.6      -
 7   false  005:02.7      -
 8   false  005:03.0      -
 9   false  005:03.1      -
10   false  005:03.2      -
11   false  005:03.3      -
12   false  005:03.4      -
13   false  005:03.5      -
14   false  005:03.6      -
15   false  005:03.7      -
```

16. 为 VM 中检测到的适配器安装 Marvell 驱动程序。为主机 OS 使用可从供应商处获得的最新驱动程序 （请勿使用内建驱动程序）。主机和 VM 上必须安装相同的驱动程序版本。
17. 打开 VM 的电源，然后发出 **`ifconfig -a`** 命令以验证添加的网络接口是否已列出。
18. 根据需要，在 VM 中添加更多 VF。

13 使用 RDMA 的 NVMe-oF 配置

结构上的非易失性存储器表示 (NVMe-oF) 允许使用到 PCIe 的备用通路，以扩展 NVMe 主机设备和 NVMe 存储驱动器或子系统可以连接的距离。NVMe-oF 定义了一个通用体系结构，它支持一系列存储网络结构，用于存储网络结构之上的 NVMe 块存储协议。该体系结构包括在存储系统中启用前端接口，扩展到大量 NVMe 设备，以及在数据中心内延长 NVMe 设备和 NVMe 子系统可访问的距离。

本章介绍的 NVMe-oF 配置步骤和选项适用于基于以太网的 RDMA 协议，包括 RoCE 和 iWARP。包含 RDMA 的 NVMe-oF 的开发由 NVMe 组织的技术小组定义。

本章演示如何在针对简单的网络上配置 NVMe-oF。示例网络 包括以下内容：

- 两台服务器：启动器服务器和目标服务器。目标服务器配备 PCIe SSD 驱动器。
- 操作系统：RHEL 7.6 及更高版本， RHEL 8.*x* 及更高版本， SLES 15.x 及更高版本
- 两个适配器：每个服务器都安装一个 41*xxx* 系列适配器。每个端口都可以独立配置为使用 RoCE、 RoCEv2 或 iWARP 作为运行 NVMe-oF 的 RDMA 协议。
- 对于 RoCE 和 RoCEv2，配置用于数据中心桥接 (DCB)、相关服务质量 (QoS) 策略和 vLAN 以承载 NVMe-oF 的 RoCE/RoCEv2 DCB 流量类别优先级的可选交换机。当 NVMe-oF 在使用 iWARP 时，不需要交换机。

图 13-1 图示了示例网络。

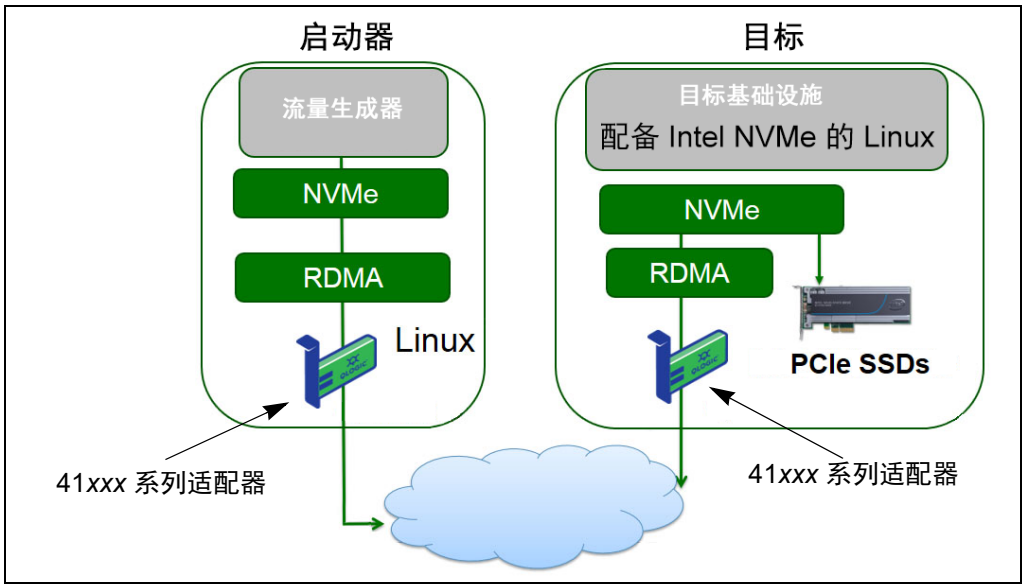


图 13-1. NVMe-oF 网络

NVMe-oF 配置过程涵盖以下程序：

- 在两台服务器上安装设备驱动程序
- 配置目标服务器
- 配置启动器服务器
- 预处理目标服务器
- 测试 NVMe-oF 设备
- 优化性能

在两台服务器上安装设备驱动程序

安装操作系统 (SLES 12 SP3) 后，请在两台服务器上安装设备驱动程序。要将内核升级到最新的 Linux 上游内核，请转至：

https://www.kernel.org/pub/linux/kernel/v4.x/

1. 按照 README 中的所有安装说明安装并加载最新的 FastLinQ 驱动程序（qed、 qede、 libqedr/qedr）。
2. （可选）如果您升级了 OS 内核，则必须重新安装并加载最新的驱动程序，如下所示：
   a. 按照 README 中的所有安装说明安装最新的 FastLinQ 固件
   b. 通过发出以下命令来安装 OS RDMA 支持应用程序和库：

```
# yum groupinstall "Infiniband Support"
# yum install tcl-devel libibverbs-devel libnl-devel
glib2-devel libudev-devel lsscsi perftest
# yum install gcc make git ctags ncurses ncurses-devel
openssl* openssl-devel elfutils-libelf-devel*
```

   c. 要确保 NVMe OFED 支持位于所选的 OS 内核中，请发出以下命令：

```
make menuconfig
```

   d. 在 **Device Drivers** （设备驱动程序）下，确保已启用以下设置 （设置为 **M**）：

```
NVM Express block devices
NVM Express over Fabrics RDMA host driver
NVMe Target support
NVMe over Fabrics RDMA target support
```

   e. （可选）如果 **Device Drivers** （设备驱动程序）选项尚不存在，请通过发出以下命令来重建内核：

```
# make
# make modules
# make modules_install
# make install
```

   f. 如果对内核进行了更改，请重新引导到该新的 OS 内核。有关如何设置默认引导内核的说明，请转至：

   https://wiki.centos.org/HowTos/Grub2

3. 启用并启动 RDMA 服务，如下所示：

```
# systemctl enable rdma.service
# systemctl start rdma.service
```

忽略 `RDMA Service Failed` 错误。qedr 所需的所有 OFED 模块已经加载。

配置目标服务器

重新引导过程后配置目标服务器。服务器运行后，不重新引导无法更改配置。如果您在使用启动脚本来配置目标服务器，请考虑根据需要暂停脚本（使用 `wait` 命令或类似的命令），以确保每个命令在执行下一个命令之前完成。

配置目标服务：

1. 加载目标模块。每次服务器重新引导后发出以下命令：

```
# modprobe qedr
# modprobe nvmet; modprobe nvmet-rdma
# lsmod | grep nvme（确认模块已加载）
```

2. 使用由 `<nvme-subsystem-name>` 指示的名称创建目标子系统 NVMe 限定名称 (NQN)。使用 NVMe-oF 规范；例如 `nqn.<YEAR>-<Month>.org.<your-company>`。

```
# mkdir /sys/kernel/config/nvmet/subsystems/<nvme-subsystem-name>
# cd /sys/kernel/config/nvmet/subsystems/<nvme-subsystem-name>
```

3. 根据需要为其他 NVMe 设备创建多个唯一的 NQN。
4. 设置目标参数，如 表 13-1 中所列。

表 13-1. 目标参数

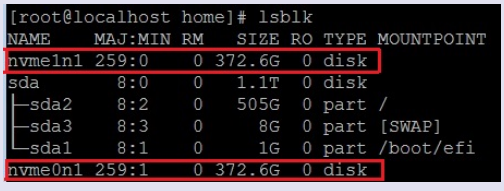
命令	说明
`# echo 1 > attr_allow_any_host`	允许连接任何主机。
`# mkdir namespaces/1`	创建命名空间。
`# echo -n /dev/nvme0n1 >namespaces/1/device_path`	设置 NVMe 设备路径。NVMe 设备路径可能因系统而异。使用 `lsblk` 命令检查设备路径。该系统有两个 NVMe 设备：`nvme0n1` 和 `nvme1n1`。 

表 13-1. 目标参数 (续)

命令	说明
`# echo 1 > namespaces/1/enable`	启用命名空间。
`# mkdir /sys/kernel/config/nvmet/ports/1` `# cd /sys/kernel/config/nvmet/ports/1`	创建 NVMe 端口 1。
`# echo 1.1.1.1 > addr_traddr`	设置相同的 IP 地址。例如， 1.1.1.1 是 41*xxx* 系列适配器的目标端口的 IP 地址。
`# echo rdma > addr_trtype`	设置传输类型 RDMA。
`# echo 4420 > addr_trsvcid`	设置 RDMA 端口号。 NVMe-oF 的套接字端口号通常为 4420。但是，如果在整个配置中始终使用端口号，则可以使用任何端口号。
`# echo ipv4 > addr_adrfam`	设置 IP 地址类型。

5. 创建一个符号链接 (symlink) 到新创建的 NQN 子系统：

```
# ln -s /sys/kernel/config/nvmet/subsystems/
nvme-subsystem-name subsystems/nvme-subsystem-name
```

6. 确认 NVMe 目标是否正在侦听端口，如下所示：

```
# dmesg | grep nvmet_rdma
[ 8769.470043] nvmet_rdma: enabling port 1 (1.1.1.1:4420)
```

配置启动器服务器

在重新引导过程之后必须配置启动器服务器。服务器运行后，不重新引导无法更改配置。如果您使用启动脚本来配置启动器服务器，请考虑根据需要暂停脚本 （使用 `wait` 命令或类似的命令），以确保每一条命令在执行下一条命令之前完成。

要配置启动器服务器：

1. 加载 NVMe 模块。每次服务器重新引导后发出这些命令：

```
# modprobe qedr
# modprobe nvme-rdma
```

2. 下载、编译并安装 `nvme-cli` 启动器公用程序。在第一次配置时发出这些命令 — 每次重新引导后不需要发出这些命令。

```
# git clone https://github.com/linux-nvme/nvme-cli.git
# cd nvme-cli
# make && make install
```

3. 验证安装版本如下：

```
# nvme version
```

4. 按照以下方式发现 NVMe-oF 目标：

```
# nvme discover -t rdma -a 1.1.1.1 -s 1023
```

记下已发现目标 (图 13-2) 的子系统 NQN (`subnqn`) 用于步骤 5。

```
[root@localhost home]# nvme discover -t rdma -a 1.1.1.1 -s 1023

Discovery Log Number of Records 1, Generation counter 1
=====Discovery Log Entry 0======
trtype:  rdma
adrfam:  ipv4
subtype: nvme subsystem
treq:    not specified
portid:  1
trsvcid: 1023

subnqn:  nvme-qlogic-tgt1
traddr:  1.1.1.1

rdma_prtype: not specified
rdma_qptype: connected
rdma_cms:    rdma-cm
rdma_pkey: 0x0000
```

图 13-2. 子系统 NQN

5. 使用 NQN 连接到发现的 NVMe-oF 目标 （`nvme-qlogic-tgt1`）。每次服务器重新引导后发出以下命令。例如：

```
# nvme connect -t rdma -n nvme-qlogic-tgt1 -a 1.1.1.1 -s 1023
```

6. 按照以下方式确认与 NVMe-oF 设备的 NVMe-oF 目标连接：

```
# dmesg | grep nvme
# lsblk
# list nvme
```

图 13-3 显示一个示例。

```
[root@localhost home] #dmesg | grep nvme
[  233.645554] nvme nvme0: new ctrl: NQN “nvme-qlogic-tgt1”, addr 1.1.1.1:1023
[root@localhost home] # lsblk
NAME      MAJ:MIN  RM     SIZE  RO TYPE  MOUNTPOINT
sdb         8:0     0     1.1T  0  disk
├─sdb2      8:2     0   493.2G  0  part  /
└─sdb3      8:3     0       8G  0  part  [SWAP]
└─sdb1      8:1     0       1G  0  part  /boot/efi
nvme0n1   259:0     0   372.6G  0  disk
[root@localhost home] # nvme list
Node            SN                   Model  Namespace   Usage                      Format           FW Rev
-------------   ------------------   -----  ----------- -------------------------  ---------------  -------
/dev/nvme0n1    7a591f3ec788a367     Linux  1            1.60  TB /   1.60  TB  512   B +  0 B  4.13.8
```

图 13-3. 确认 NVMe-oF 连接

预处理目标服务器

开箱即用的 NVMe 目标服务器显示的性能高于预期。在运行基准测试之前，需要*预填充*或*预处理*目标服务器。

要预处理目标服务器：

1. 使用供应商特定工具安全擦除目标服务器 （类似于格式化）。此测试示例使用 Intel NVMe SSD 设备，该设备需要以下链接提供的 Intel 数据中心工具：

   https://downloadcenter.intel.com/download/23931/Intel-Solid-State-Drive-Data-Center-Tool

2. 使用数据对目标服务器 (`nvme0n1`) 进行预处理，以确保填充所有可用内存。此示例使用 “DD” 磁盘公用程序：

   ```
   # dd if=/dev/zero bs=1024k of=/dev/nvme0n1
   ```

测试 NVMe-oF 设备

将目标服务器上的本地 NVMe 设备的延迟与启动器服务器上的 NVMe-oF 设备的延迟进行比较，以显示 NVMe 添加到系统的延迟。

要测试 NVMe-oF 设备：

1. 更新存储库 (Repo) 源并通过发出以下命令在目标服务器和启动器服务器上安装灵活输入 / 输出 (FIO) 基准测试公用程序：

   ```
   # yum install epel-release
   ```

```
# yum install fio
```

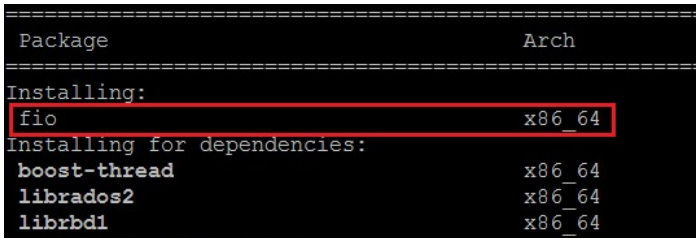


图 13-4. FIO 公用程序安装

2. 运行 FIO 公用程序来测量启动器 NVMe-oF 设备的延迟。发出以下命令：

```
# fio --filename=/dev/nvme0n1 --direct=1 --time_based
--rw=randread --refill_buffers --norandommap --randrepeat=0
--ioengine=libaio --bs=4k --iodepth=1 --numjobs=1
--runtime=60 --group_reporting --name=temp.out
```

FIO 报告两种延迟类型：提交和完成。提交延迟 (slat) 测量应用程序到内核的延迟。完成延迟 (clat) 测量端到端的内核延迟。业界公认的方法是在第 99.00 范围内读取 *clat 百分位数*。

在本示例中，启动器设备 NVMe-oF 的延迟为 30µsec。

3. 运行 FIO 以测量目标服务器上本地 NVMe 设备的延迟。发出以下命令：

```
# fio --filename=/dev/nvme0n1 --direct=1 --time_based
--rw=randread --refill_buffers --norandommap --randrepeat=0
--ioengine=libaio --bs=4k --iodepth=1 --numjobs=1
--runtime=60 --group_reporting --name=temp.out
```

在本示例中，目标 NVMe 设备的延迟为 8µsec。使用 NVMe-oF 产生的总延迟是启动器设备 NVMe-oF 延迟 (30µsec) 与目标设备 NVMe-oF 延迟 (8µsec) 之间的差异，或 22µsec。

4. 运行 FIO 以测量目标服务器上本地 NVMe 设备的带宽。发出以下命令：

```
fio --verify=crc32 --do_verify=1 --bs=8k --numjobs=1
--iodepth=32 --loops=1 --ioengine=libaio --direct=1
--invalidate=1 --fsync_on_close=1 --randrepeat=1
--norandommap --time_based --runtime=60
--filename=/dev/nvme0n1 --name=Write-BW-to-NVMe-Device
--rw=randwrite
```

其中 `--rw` 可以是只读的 `randread` 、只写的 `randwrite` 或可读取可写入的 `randrw`。

优化性能

要优化启动器服务器和目标服务器的性能：

1. 配置以下系统 BIOS 设置：
   - ❑ Power Profiles ='Max Performance' 或等价对象
   - ❑ ALL C-States = Disabled （已禁用）
   - ❑ Hyperthreading = Disabled （已禁用）
2. 通过编辑 `grub` 文件 (`/etc/default/grub`) 配置 Linux 内核参数。
   a. 将参数添加到行 `GRUB_CMDLINE_LINUX` 的结尾：

   ```
   GRUB_CMDLINE_LINUX="nosoftlockup intel_idle.max_cstate=0
   processor.max_cstate=1 mce=ignore_ce idle=poll"
   ```

   b. 保存 `grub` 文件。
   c. 重建 `grub` 文件。
      - ■ 要为旧版 BIOS 引导重建 `grub` 文件，请发出以下命令：

      ```
      # grub2-mkconfig -o /boot/grub2/grub.cfg （旧版 BIOS 引导）
      ```

      - ■ 要为 EFI 引导重建 `grub` 文件，请发出以下命令：

      ```
      # grub2-mkconfig -o /boot/efi/EFI/<os>/grub.cfg （EFI 引导）
      ```

   d. 重新引导服务器实施更改。
3. 为所有 41*xxx* 系列适配器 设置 IRQ 关联。`multi_rss-affin.sh` 文件是 第 242 页上 “.IRQ 关联 (multi_rss-affin.sh)” 中列出的一个脚本文件。

   ```
   # systemctl stop irqbalance
   # ./multi_rss-affin.sh eth1
   ```

   注

   此脚本的不同版本 `qedr_affin.sh` 位于 `\add-ons\performance\roce` 目录中的 41*xxx* Linux 源代码包中。有关 IRQ 关联设置的说明，请参阅该目录中的 `multiple_irqs.txt` 文件。
4. 设置 CPU 频率。`cpufreq.sh` 文件是 第 243 页上 “CPU 频率 (cpufreq.sh)” 中列出的一个脚本。

   ```
   # ./cpufreq.sh
   ```

以下各节列出了用于 步骤 3 和 4 的脚本。

.IRQ 关联 (multi_rss-affin.sh)

以下脚本设置 IRQ 关联。

```
#!/bin/bash
#RSS affinity setup script
#input: the device name (ethX)
#OFFSET=0   0/1   0/1/2   0/1/2/3
#FACTOR=1    2      3       4
OFFSET=0
FACTOR=1
LASTCPU='cat /proc/cpuinfo | grep processor | tail -n1 | cut -d":" -f2'
MAXCPUID='echo 2 $LASTCPU ^  p | dc'
OFFSET='echo 2 $OFFSET ^  p | dc'
FACTOR='echo 2 $FACTOR ^  p | dc'
CPUID=1

for eth in $*; do

NUM='grep $eth /proc/interrupts | wc -l'
NUM_FP=$((${NUM}))

INT='grep -m 1 $eth /proc/interrupts | cut -d ":" -f 1'

echo  "$eth: ${NUM} (${NUM_FP} fast path) starting irq ${INT}"

CPUID=$((CPUID*OFFSET))
for ((A=1; A<=${NUM_FP}; A=${A}+1)) ; do
INT='grep -m $A $eth /proc/interrupts | tail -1  | cut -d ":" -f 1'
SMP='echo $CPUID 16 o p | dc'
echo ${INT} smp affinity set to ${SMP}
echo  $((${SMP})) > /proc/irq/$((${INT}))/smp_affinity
CPUID=$((CPUID*FACTOR))
if [ ${CPUID} -gt ${MAXCPUID} ]; then
CPUID=1
CPUID=$((CPUID*OFFSET))
fi
done
done
```

CPU 频率 (cpufreq.sh)

以下脚本设置 CPU 频率。

```
#Usage "./nameofscript.sh"
grep -E '^model name|^cpu MHz' /proc/cpuinfo
cat /sys/devices/system/cpu/cpu0/cpufreq/scaling_governor
for CPUFREQ in /sys/devices/system/cpu/cpu*/cpufreq/scaling_governor; do [ -f
$CPUFREQ ] || continue; echo -n performance > $CPUFREQ; done
cat /sys/devices/system/cpu/cpu0/cpufreq/scaling_governor
```

要配置网络或内存设置：

```
sysctl -w net.ipv4.tcp_mem="16777216 16777216 16777216"
sysctl -w net.ipv4.tcp_wmem="4096 65536 16777216"
sysctl -w net.ipv4.tcp_rmem="4096 87380 16777216"
sysctl -w net.core.wmem_max=16777216
sysctl -w net.core.rmem_max=16777216
sysctl -w net.core.wmem_default=16777216
sysctl -w net.core.rmem_default=16777216
sysctl -w net.core.optmem_max=16777216
sysctl -w net.ipv4.tcp_low_latency=1
sysctl -w net.ipv4.tcp_timestamps=0
sysctl -w net.ipv4.tcp_sack=1
sysctl -w net.ipv4.tcp_window_scaling=0
sysctl -w net.ipv4.tcp_adv_win_scale=1
```

注

以下命令仅适用于启动器服务器。

```
# echo 0 > /sys/block/nvme0n1/queue/add_random
# echo 2 > /sys/block/nvme0n1/queue/nomerges
```

14 VXLAN 配置

本章提供以下内容的说明：

在 Linux 上配置 VXLAN

要在 Linux 上配置 VXLAN：

1. 下载、提取并配置 openvswitch （开放虚拟交换机， OVS）原始码。

   a. 从以下地址下载正确的 openvswitch 版本：

   http://www.openvswitch.org/download/

   b. 通过导航至下载 openvswitch 版本的目录，并发出以下命令来提取原始码：

   `./configure; make; make install` （编译）

   c. 通过发出以下命令配置 openvswitch：

```
modprobe -v openvswitch
export PATH=$PATH:/usr/local/share/openvswitch/scripts
ovs-ctl start
ovs-ctl status
```

   当运行 `ovs-ctl` 状态时， `ovsdb-server` 和 `ovs-vswitchd` 应与 `pid` 一起运行。例如：

```
[root@localhost openvswitch-2.11.1]# ovs-ctl status
ovsdb-server is running with pid 8479
ovs-vswitchd is running with pid 8496
```

2. 创建桥接。

   a. 要配置主机 1，请发出以下命令：

```
ovs-vsctl add-br br0
ovs-vsctl add-br br1
ovs-vsctl add-port br0 eth0
ifconfig eth0 0 && ifconfig br0 192.168.1.10 netmask 255.255.255.0
route add default gw 192.168.1.1 br0
ifconfig br1 10.1.2.10 netmask 255.255.255.0
ovs-vsctl add-port br1 vx1 -- set interface vx1 type=vxlan
options:remote_ip=192.168.1.11 （对等机 IP 地址）
```

   b. 要配置主机 2，请发出以下命令：

```
ovs-vsctl add-br br0
ovs-vsctl add-br br1
ovs-vsctl add-port br0 eth0
ifconfig eth0 0 && ifconfig br0 192.168.1.11 netmask 255.255.255.0
route add default gw 192.168.1.1 br0
ifconfig br1 10.1.2.11 netmask 255.255.255.0
ovs-vsctl add-port br1 vx1 -- set interface vx1 type=vxlan options:
remote_ip=192.168.1.10
```

3. 验证配置。

   使用 iperf 运行主机和对等机之间的流量。请确保防火墙和 iptables 各自停止并清除。

4. 配置桥接作为 VM 的直通，然后检查 VM 与对等机之间的连接。

   a. 通过虚拟交换机管理器创建 VM。

   b. 由于没有选项可通过虚拟交换机管理器连接桥接 `br1`，请执行以下操作更改 xml 文件：

      发出以下命令：

      ```
      command:  virsh edit vm1
      ```

      添加以下代码：

      ```
      <interface type='bridge'>
      <source bridge='br1'/>
      <virtualport type='openvswitch'>
      <parameters/>
      </virtualport>
      <model type='virtio'/>
      </interface>
      ```

   c. 启动 VM 并检查 `br1` 接口。

      请确保 br1 位于 OS 中。`br1` 接口命名为 eth0、 ens7 ；通过网络设备文件手动配置静态 Ip，并将相同的子网 IP 分配给对等机 （主机 2 虚拟机）。

      运行对等机和 VM 之间的流量。

注

您可使用此程序通过 OVS 检测其它隧道，例如通用网络虚拟化封装 (GENEVE) 和通用路由封装 (GRE)。

如果您不想使用 OVS，您可通过旧版桥接选项 `brctl` 继续。

在 VMware 中配置 VXLAN

要在 VMware 中配置 VXLAN，请遵循以下网址的指示：

https://docs.vmware.com/en/VMware-NSX-Data-Center-for-vSphere/6.3/com.vmware.nsx.cross-vcenter-install.doc/GUID-49BAECC2-B800-4670-AD8C-A5292ED6BC19.html

https://www.vmware.com/content/dam/digitalmarketing/vmware/en/pdf/products/nsx/vmw-nsx-network-virtualization-design-guide.pdf

https://pubs.vmware.com/nsx-63/topic/com.vmware.nsx.troubleshooting.doc/GUID-EA1DB524-DD2E-4157-956E-F36BDD20CDB2.html

https://communities.vmware.com/api/core/v3/attachments/124957/data

在 Windows Server 2016 中配置 VXLAN

Windows Server 2016 中的 VXLAN 配置包括：

- 在适配器上启用 VXLAN 卸载
- 部署软件定义网络 (SDN)

在适配器上启用 VXLAN 卸载

要在适配器上启用 VXLAN 卸载：

1. 打开微型端口属性，然后单击 **Advanced** （高级）选项卡。
2. 在适配器属性的 Advanced （高级）页面 （图 14-1）的 **Property** （属性）下，选择 **VXLAN Encapsulated Task Offload** （VXLAN 封装任务卸载）。

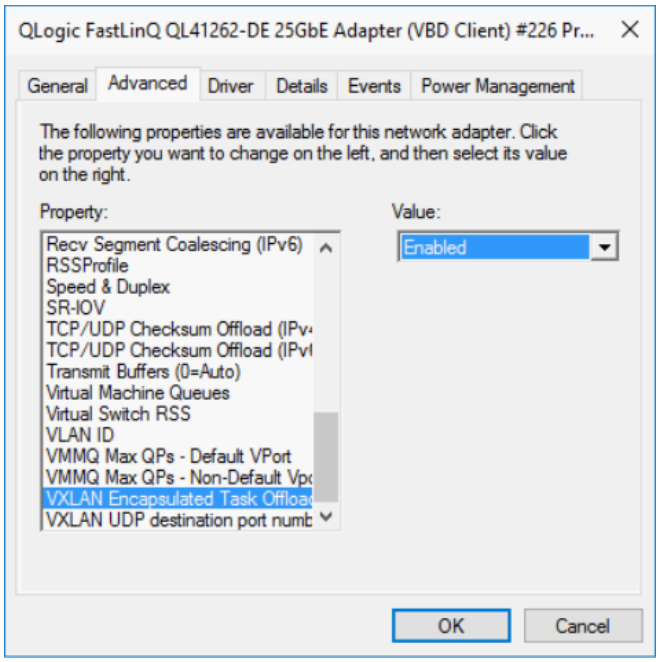


图 14-1. 高级属性：启用 VXLAN

3. 将 **Value** （值）设置为 **Enabled** （已启用）。
4. 单击 **OK** （确定）。

部署软件定义网络 (SDN)

要利用虚拟机上的 VXLAN 封装任务卸载，必须部署使用 Microsoft 网络控制器的软件定义网络 (SDN) 堆栈。

有关更多详细信息，请参阅有关软件定义网络的以下 Microsoft TechNet 链接：

https://technet.microsoft.com/en-us/windows-server-docs/networking/sdn/software-defined-networking--sdn-

15 Windows Server 2016

本章提供有关 Windows Server 2016 的以下信息：

- 使用 Hyper-V 配置 RoCE 接口
- 第 255 页上 “Switch Embedded Teaming 上的 RoCE”
- 第 256 页上 “ 为 RoCE 配置 QoS”
- 第 265 页上 “ 配置 VMMQ”
- 第 269 页上 “ 配置 Storage Spaces Direct”

使用 Hyper-V 配置 RoCE 接口

在 Windows Server 2016 中，使用网络直接内核提供商接口 (NDKPI) 模式 2 的 Hyper-V 主机虚拟网络适配器 （主机虚拟 NIC）支持 RDMA。

注

Hyper-V 上的 RoCE 需要 DCBX。要配置 DCBX：

- 通过 HII 配置 （请参阅第 128 页上 “ 准备适配器 ”）。
- 使用 QoS 配置 （请参阅第 256 页上 “ 为 RoCE 配置 QoS”）。

本节中的 RoCE 配置步骤包括：

- 创建带 RDMA NIC 的 Hyper-V 虚拟交换机
- 将 vLAN ID 添加到主机虚拟 NIC
- 验证 RoCE 是否启用
- 添加主机虚拟 NIC （虚拟端口）
- 映射 SMB 驱动器和运行 RoCE 流量

创建带 RDMA NIC 的 Hyper-V 虚拟交换机

按照本节中的步骤操作以创建 Hyper-V 虚拟交换机，然后在主机 VNIC 中启用 RDMA。

要创建带 RDMA 虚拟 NIC 的 Hyper-V 虚拟交换机：

1. 在所有物理接口上，将 **NetworkDirect Functionality** （网络直接功能）参数的值设为 **Enabled** （已启用）。
2. 启动 Hyper-V 管理器。
3. 单击 **Virtual Switch Manager** （虚拟交换机管理器）（请参阅图 15-1）。

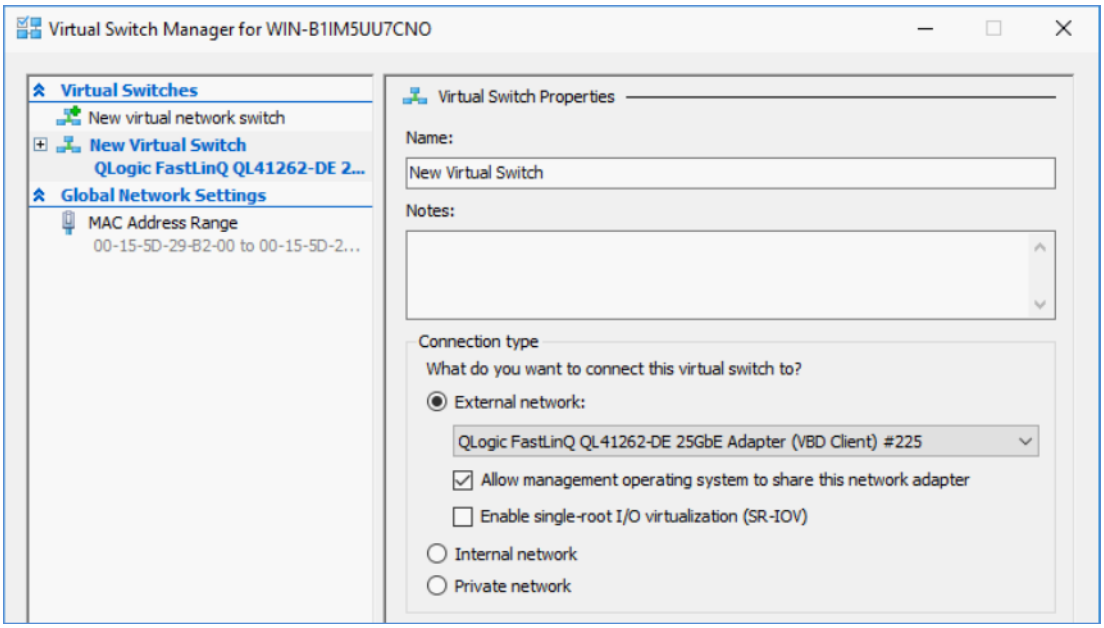


图 15-1. 在主机虚拟 NIC 中启用 RDMA

4. 创建一个虚拟交换机。
5. 选择 **Allow management operating system to share this network adapter** （允许管理操作系统以共享此网络适配器）复选框。

在 Windows Server 2016 中，将在主机虚拟 NIC 中添加一个新参数：Network Direct (RDMA)。

要在主机虚拟 NIC 中启用 RDMA：

1. 打开 Hyper-V Virtual Ethernet Adapter Properties （Hyper-V 虚拟以太网适配器属性）窗口。
2. 单击 **Advanced** （高级）选项卡。

3. 在 Advanced （高级）页面 （图 15-2）上：
   a. 在 **Property** （属性）下，选择 **Network Direct (RDMA)**。
   b. 在 **Value** （值）下，选择 **Enabled** （启用）。
   c. 单击 **OK** （确定）。

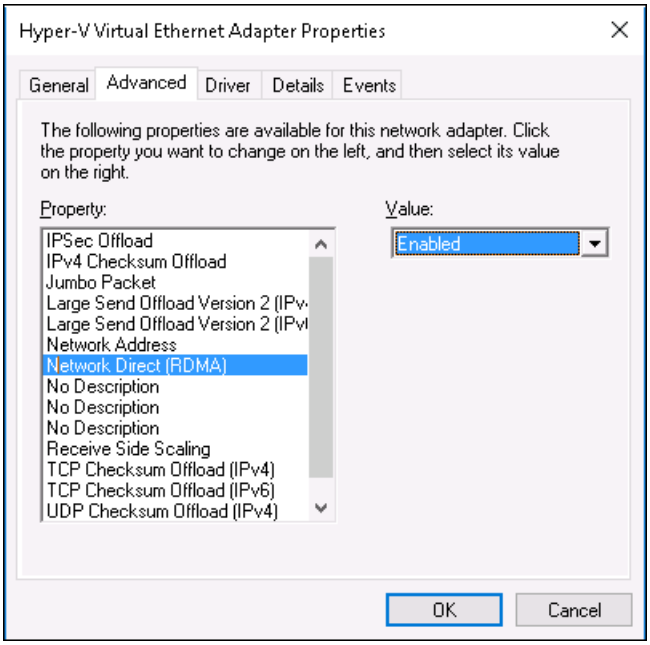


图 15-2. Hyper-V 虚拟以太网适配器属性

4. 要通过 PowerShell 启用 RDMA，请发出以下 Windows PowerShell 命令：

```
PS C:\Users\Administrator> Enable-NetAdapterRdma "vEthernet
(New Virtual Switch)"
PS C:\Users\Administrator>
```

将 vLAN ID 添加到主机虚拟 NIC

要将 vLAN ID 添加到主机虚拟 NIC：

1. 要查找主机虚拟 NIC 名称，请发出以下 Windows PowerShell 命令：

```
PS C:\Users\Administrator> Get-VMNetworkAdapter -ManagementOS
```

图 15-3 显示命令输出。

```
PS C:\Users\Administrator> Get-VMNetworkAdapter -ManagementOS

Name                IsManagementOs VMName SwitchName         MacAddress   Status IPAddresses
----                -------------- ------ ----------         ----------   ------ -----------
New Virtual Switch True                   New Virtual Switch 000E1EC41F0B {Ok}
```

图 15-3. Windows PowerShell 命令：Get-VMNetworkAdapter

2. 要设置主机虚拟 NIC 的 vLAN ID，请发出以下 Windows PowerShell 命令：

```
PS C:\Users\Administrator> Set-VMNetworkAdaptervlan
-VMNetworkAdapterName "New Virtual Switch" -VlanId 5 -Access
-ManagementOS
```

注

请注意关于将 vLAN ID 添加到主机虚拟 NIC 的以下事项：

- 必须为主机虚拟 NIC 分配一个 vLAN ID。必须为交换机上端口分配相同的 vLAN ID。
- 将主机虚拟 NIC 用于 RoCE 时，确保 vLAN ID 没有分配给物理接口。
- 如果创建多个主机虚拟 NIC，可以为每个主机虚拟 NIC 分配不同的 vLAN。

验证 RoCE 是否启用

要验证 RoCE 是否启用：

- 发出以下 Windows PowerShell 命令：

```
Get-NetAdapterRdma
```

命令输出列出 RDMA 支持的适配器，如图 15-4 中所示。

```
PS C:\Users\Administrator> Get-NetAdapterRdma

Name                      InterfaceDescription                Enabled
----                      --------------------                -------
vEthernet (New Virtual... Hyper-V Virtual Ethernet Adapter    True
```

图 15-4. Windows PowerShell 命令：Get-NetAdapterRdma

添加主机虚拟 NIC （虚拟端口）

要添加主机虚拟 NIC：

1. 要添加主机虚拟 NIC，请发出以下命令：

   ```
   Add-VMNetworkAdapter -SwitchName "New Virtual Switch" -Name
   SMB - ManagementOS
   ```

2. 在主机虚拟 NIC 上启用 RDMA，如第 250 页上 “ 要在主机虚拟 NIC 中启用 RDMA：” 中所示。

3. 要将 vLAN ID 分配给虚拟端口，请发出以下命令：

   ```
   Set-VMNetworkAdapterVlan -VMNetworkAdapterName SMB -VlanId 5
   -Access -ManagementOS
   ```

映射 SMB 驱动器和运行 RoCE 流量

要映射 SMB 驱动器和运行 RoCE 流量：

1. 启动性能监视器 (Perfmon)。

2. 填写 Add Counters （添加计数器）对话框 （图 15-5）如下：

   a. 在 **Available counters** （可用计数器）下，选择 **RDMA Activity** （RDMA 活动）。

   b. 在 **Instances of selected object** （所选对象的实例）下，选择适配器。

   c. 单击**添加**。

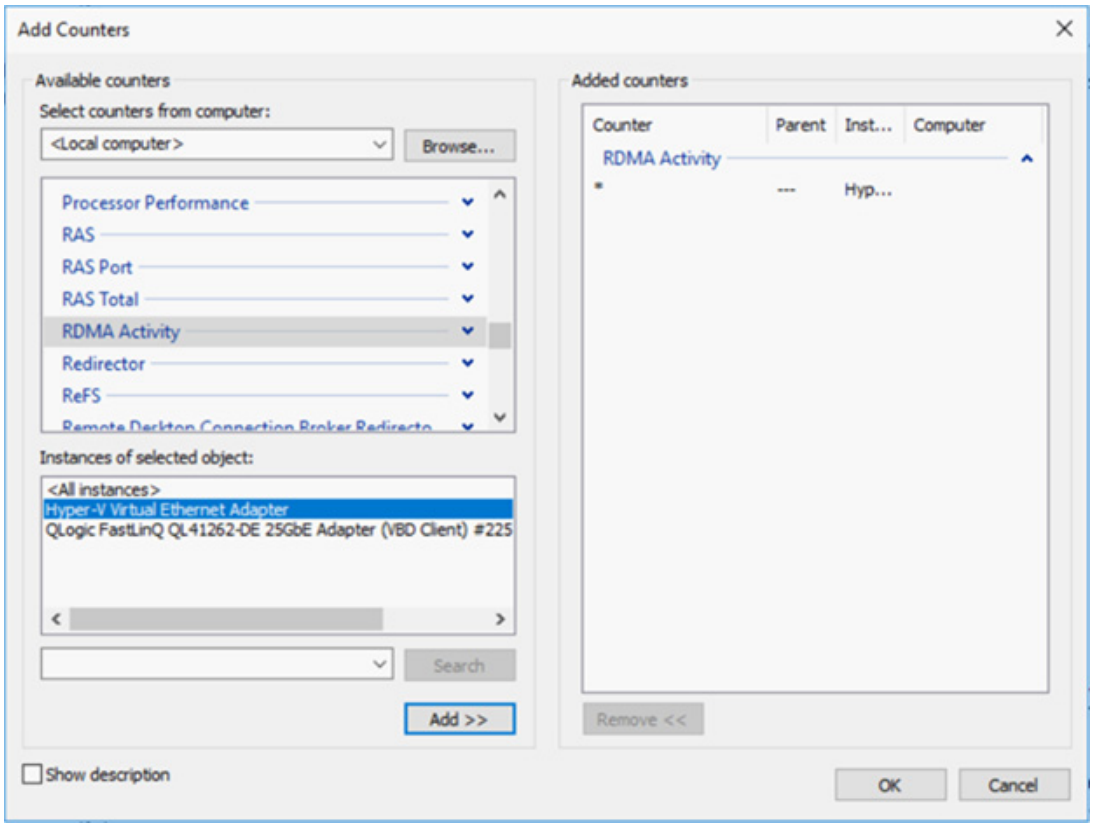


图 15-5. 添加计数器对话框

如果 RoCE 流量正在运行，计数器将显示为如图 15-6 中所示。

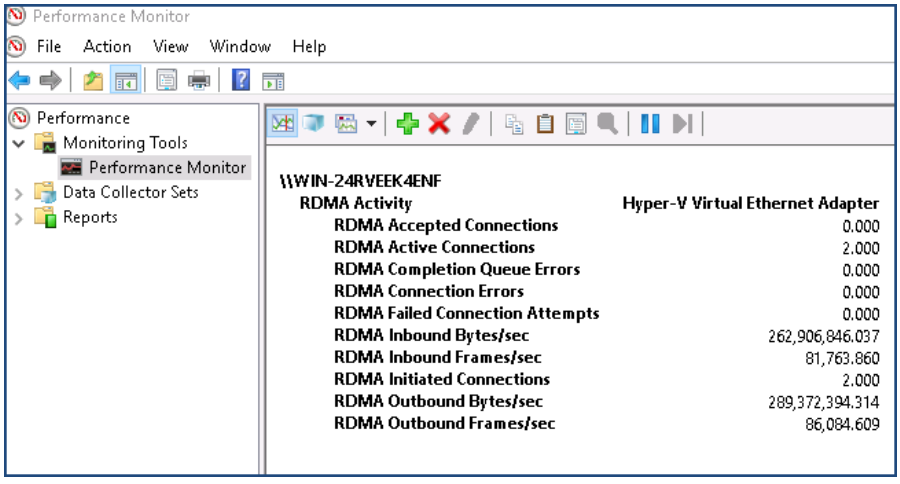


图 15-6. 显示 RoCE 流量的性能监视器

Switch Embedded Teaming 上的 RoCE

Switch Embedded Teaming (SET) 是 Microsoft 的备选 NIC 组合解决方案，可在 Windows Server 2016 Technical Preview 中包括 Hyper-V 和软件定义网络 (SDN) 堆栈的环境中使用。 SET 将受限的 NIC 组合功能集成到 Hyper-V 虚拟交换机中。

使用 SET 将一到八个物理以太网络适配器分组为一个或多个基于软件的虚拟网络适配器。如果网络适配器发生故障，这些适配器可提供快速性能和容错。 SET 成员网络适配器必须均安装在同一物理 Hyper-V 主机中方可放在一个组中。

本节中包括以下 SET 上的 RoCE 步骤：

- 创建带 SET 和 RDMA 虚拟 NIC 的 Hyper-V 虚拟交换机
- 在 SET 上启用 RDMA
- 在 SET 上分配 vLAN ID
- 在 SET 上运行 RDMA 流量

创建带 SET 和 RDMA 虚拟 NIC 的 Hyper-V 虚拟交换机

要创建带 SET 和 RDMA 虚拟 NIC 的 Hyper-V 虚拟交换机：

- 要创建 SET，请发出以下 Windows PowerShell 命令：

```
PS C:\Users\Administrator> New-VMSwitch -Name SET
-NetAdapterName "Ethernet 2","Ethernet 3"
-EnableEmbeddedTeaming $true
```

图 15-7 显示命令输出。

```
PS C:\Users\Administrator> New-VMSwitch -Name SET -NetAdapterName "Ethernet 2","Ethernet 3" -EnableEmbeddedTeaming $true

Name SwitchType NetAdapterInterfaceDescription
---- ---------- ----------------------------
SET  External   Teamed-Interface
```

图 15-7. Windows PowerShell 命令：New-VMSwitch

在 SET 上启用 RDMA

要在 SET 上启用 RDMA：

1. 要查看适配器上的 SET，请发出以下 Windows PowerShell 命令：

```
PS C:\Users\Administrator> Get-NetAdapter "vEthernet (SET)"
```

图 15-8 显示命令输出。

```
PS C:\Users\Administrator> Get-NetAdapter "vEthernet (SET)"

Name                      InterfaceDescription                    ifIndex Status       MacAddress             LinkSpeed
----                      --------------------                    ------- ------       ----------             ---------
vEthernet (SET)           Hyper-V Virtual Ethernet Adapter             46 Up           00-0E-1E-C4-04-F8       50 Gbps
```

图 15-8. Windows PowerShell 命令：Get-NetAdapter

2. 要在 SET 上启用 RDMA，请发出以下 Windows PowerShell 命令：

```
PS C:\Users\Administrator> Enable-NetAdapterRdma "vEthernet (SET)"
```

在 SET 上分配 vLAN ID

要在 SET 上分配 vLAN ID：

- 要在 SET 上分配 vLAN ID，请发出以下 Windows PowerShell 命令：

```
PS C:\Users\Administrator> Set-VMNetworkAdapterVlan -VMNetworkAdapterName "SET" -VlanId 5 -Access -ManagementOS
```

注

请注意将 vLAN ID 添加到主机虚拟 NIC 时的以下事项：

- 将主机虚拟 NIC 用于 RoCE 时，确保 vLAN ID 没有分配给物理接口。
- 如果创建多个主机虚拟 NIC，可以为每个主机虚拟 NIC 分配不同的 vLAN。

在 SET 上运行 RDMA 流量

有关在 SET 上运行 RDMA 流量的信息，请访问：

https://technet.microsoft.com/en-us/library/mt403349.aspx

为 RoCE 配置 QoS

配置服务质量 (QoS) 的两种方法包括：

- 通过在适配器上禁用 DCBX 配置 QoS
- 通过在适配器上启用 DCBX 配置 QoS

通过在适配器上禁用 DCBX 配置 QoS

通过在适配器上禁用 DCBX 配置 QoS 之前，必须完成所有使用中系统的所有配置。基于优先级的流控制 (PFC)、增强的转换服务 (ETS) 和流量类配置在交换机和服务器上必须相同。

要通过禁用 DCBX 配置 QoS：

1. 在适配器上禁用 DCBX。
2. 使用 HII，将 **RoCE Priority** （RoCE 优先级）设置为 **0**。
3. 要在主机中安装 DCB 角色，请发出以下 Windows PowerShell 命令：

   ```
   PS C:\Users\Administrators> Install-WindowsFeature
   Data-Center-Bridging
   ```

4. 要将 **DCBX Willing** （DCBX 愿意）模式设置为 **False** （假），请发出以下 Windows PowerShell 命令：

   ```
   PS C:\Users\Administrators> set-NetQosDcbxSetting -Willing 0
   ```

5. 请如下在微型端口中启用 QoS：
   a. 打开微型端口属性，然后单击 **Advanced** （高级）选项卡。
   b. 在适配器属性的 Advanced （高级）页面 （图 15-9）的 **Property** （属性）下，选择 **Quality of Service** （服务质量），然后将值设置为 **Enabled** （已启用）。
   c. 单击 **OK** （确定）。

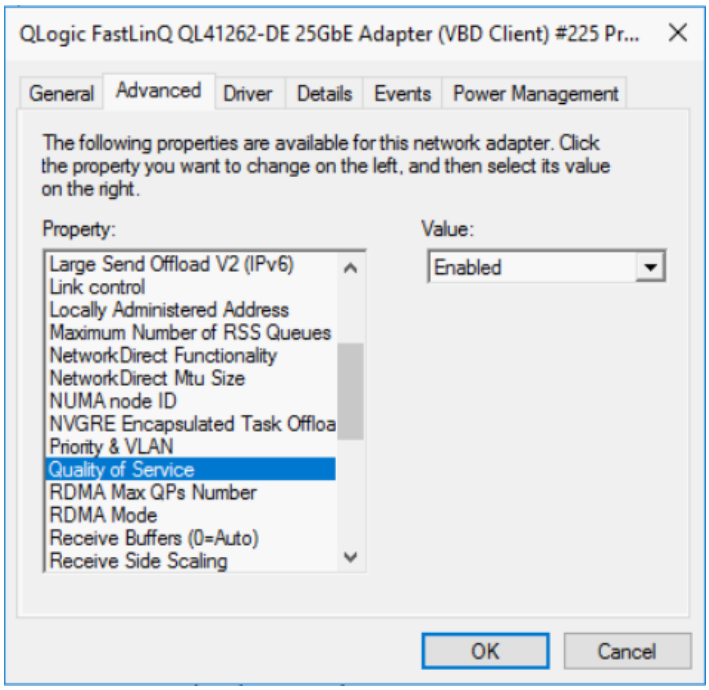


图 15-9. 高级属性：启用 QoS

6. 请如下所示将 vLAN ID 分配给接口：

   a. 打开微型端口属性，然后单击 **Advanced**（高级）选项卡。

   b. 在适配器属性的 Advanced（高级）页面（图 15-10）的 **Property**（属性）下，选择 **VLAN ID**，然后设置值。

   c. 单击 **OK**（确定）。

注

优先级流控制 (PFC) 需要上述步骤。

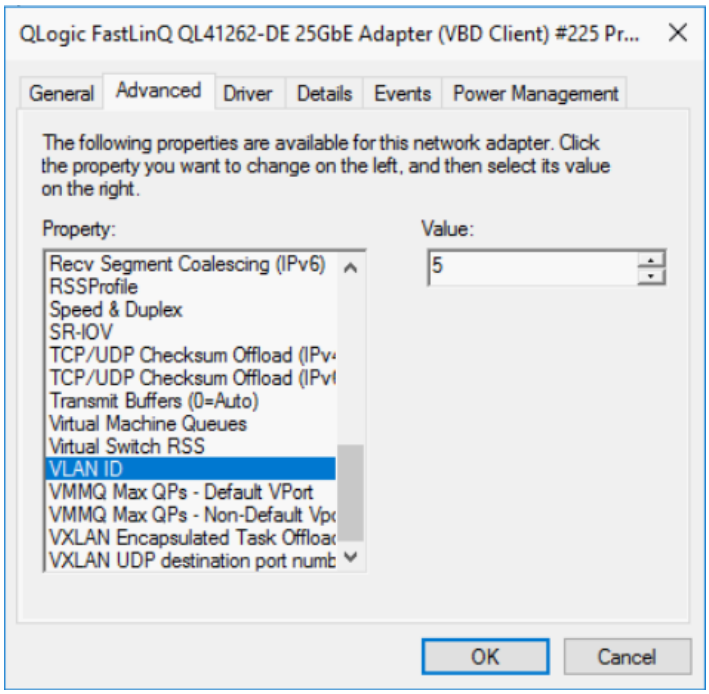


图 15-10. 高级属性：设置 VLAN ID

7. 要基于特定优先级为 RoCE 启用 PFC，请发出以下命令：

```
PS C:\Users\Administrators> Enable-NetQoSFlowControl
-Priority 5
```

注

如果通过 Hyper-V 配置 RoCE，请勿将 vLAN ID 分配给物理接口。

8. 要基于任何其他优先级禁用优先级流控制，请发出以下命令：

```
PS C:\Users\Administrator> Disable-NetQosFlowControl 0,1,2,3,4,6,7
PS C:\Users\Administrator> Get-NetQosFlowControl
Priority   Enabled    PolicySet        IfIndex IfAlias
--------   -------    ---------        ------- -------
0          False      Global
1          False      Global
2          False      Global
3          False      Global
4          False      Global
```

```
5           True        Global
6           False       Global
7           False       Global
```

9. 要为每种类型的流量配置 QoS 和分配相关的优先级，请发出以下命令（其中，优先级 5 标记用于 RoCE 而优先级 0 标记用于 TCP）：

```
PS C:\Users\Administrators> New-NetQosPolicy "SMB"
-NetDirectPortMatchCondition 445 -PriorityValue8021Action 5 -PolicyStore
ActiveStore

PS C:\Users\Administrators> New-NetQosPolicy "TCP" -IPProtocolMatchCondition
TCP -PriorityValue8021Action 0 -Policystore ActiveStore

PS C:\Users\Administrator> Get-NetQosPolicy -PolicyStore activestore

Name            : tcp
Owner           : PowerShell / WMI
NetworkProfile  : All
Precedence      : 127
JobObject       :
IPProtocol      : TCP
PriorityValue   : 0

Name            : smb
Owner           : PowerShell / WMI
NetworkProfile  : All
Precedence      : 127
JobObject       :
NetDirectPort   : 445
PriorityValue   : 5
```

10. 要为之前步骤中定义的所有流量类配置 ETS，请发出以下命令：

```
PS C:\Users\Administrators> New-NetQosTrafficClass -name "RDMA class"
-priority 5 -bandwidthPercentage 50 -Algorithm ETS

PS C:\Users\Administrators> New-NetQosTrafficClass -name "TCP class" -priority
0 -bandwidthPercentage 30 -Algorithm ETS

PS C:\Users\Administrator> Get-NetQosTrafficClass
```

```
Name        Algorithm Bandwidth(%) Priority    PolicySet       IfIndex IfAlias
----        --------- ------------ --------    ---------       ------- -------
[Default]    ETS       20           1-4,6-7     Global
RDMA class   ETS       50           5           Global
TCP class    ETS       30           0           Global
```

11. 要查看之前配置的网络适配器 QoS，请发出以下 Windows PowerShell 命令：

```
PS C:\Users\Administrator> Get-NetAdapterQos

Name                          : SLOT 4 Port 1
Enabled                       : True
Capabilities                  :                          Hardware     Current
                                                         --------     -------
                                MacSecBypass         : NotSupported NotSupported
                                DcbxSupport          : None         None
                                NumTCs(Max/ETS/PFC)  : 4/4/4        4/4/4

OperationalTrafficClasses     : TC TSA    Bandwidth Priorities
                                -- ---    --------- ----------
                                 0 ETS    20%       1-4,6-7
                                 1 ETS    50%       5
                                 2 ETS    30%       0

OperationalFlowControl        : Priority 5 Enabled
OperationalClassifications    : Protocol  Port/Type Priority
                                --------  --------- --------
                                Default             0
                                NetDirect 445       5
```

12. 创建启动脚本，以使设置在系统重新引导后保持不变。
13. 运行 RDMA 流量并验证，如第 126 页上 “RoCE 配置 ” 中所述。

通过在适配器上启用 DCBX 配置 QoS

必须在所有使用中的系统上完成所有配置。 PFC、 ETS 和流量类配置在交换机和服务器上必须相同。

要通过启用 DCBX 配置 QoS：

1. 启用 DCBX （IEEE、 CEE 或动态）。
2. 使用 HII，将 **RoCE Priority** （RoCE 优先级）设置为 **0**。

注

如果交换机无法指定 RoCE 流量，您可能需要将 **RoCE Priority**（RoCE 优先级）设为交换机使用的值。Arista® 交换机可以，但有些其他的交换机不行。

3. 要在主机中安装 DCB 角色，请发出以下 Windows PowerShell 命令：

```
PS C:\Users\Administrators> Install-WindowsFeature
Data-Center-Bridging
```

注

对于此配置，将 **DCBX Protocol**（DCBX 协议）设置为 **CEE**。

4. 要将 **DCBX Willing**（DCBX 愿意）模式设置为 **True**（真），请发出以下命令：

```
PS C:\Users\Administrators> set-NetQosDcbxSetting -Willing 1
```

5. 请如下在微型端口属性中启用 QoS：

   a. 在适配器属性的 Advanced（高级）页面（图 15-11）的 **Property**（属性）下，选择 **Quality of Service**（服务质量），然后将值设置为 **Enabled**（已启用）。

   b. 单击 **OK**（确定）。

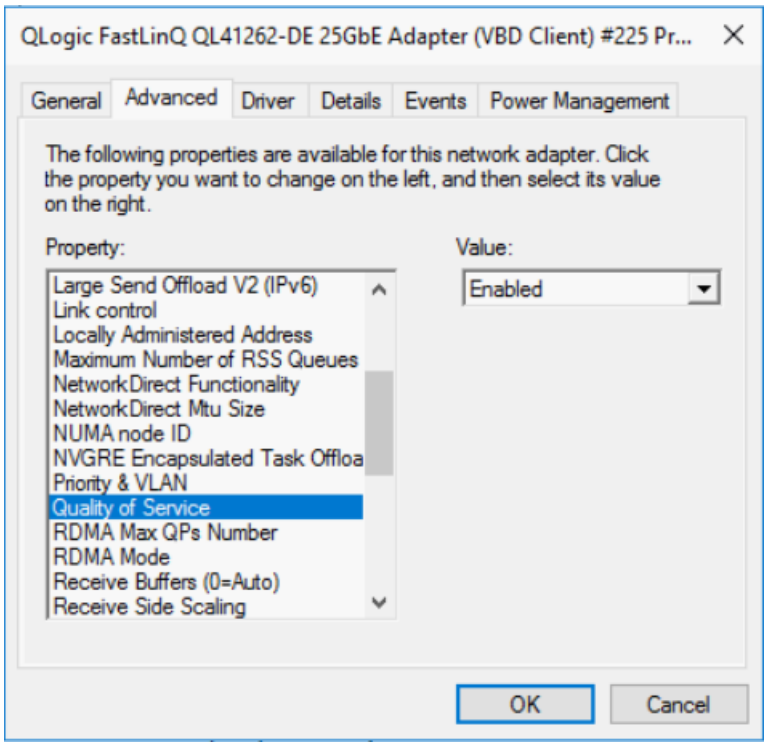


图 15-11. 高级属性：启用 QoS

6. 请如下将 vLAN ID 分配给接口 （PFC 需要）：
   a. 打开微型端口属性，然后单击 **Advanced** （高级）选项卡。
   b. 在适配器属性的 Advanced （高级）页面 （图 15-12）的 **Property** （属性）下，选择 **VLAN ID**，然后设置值。
   c. 单击 **OK** （确定）。

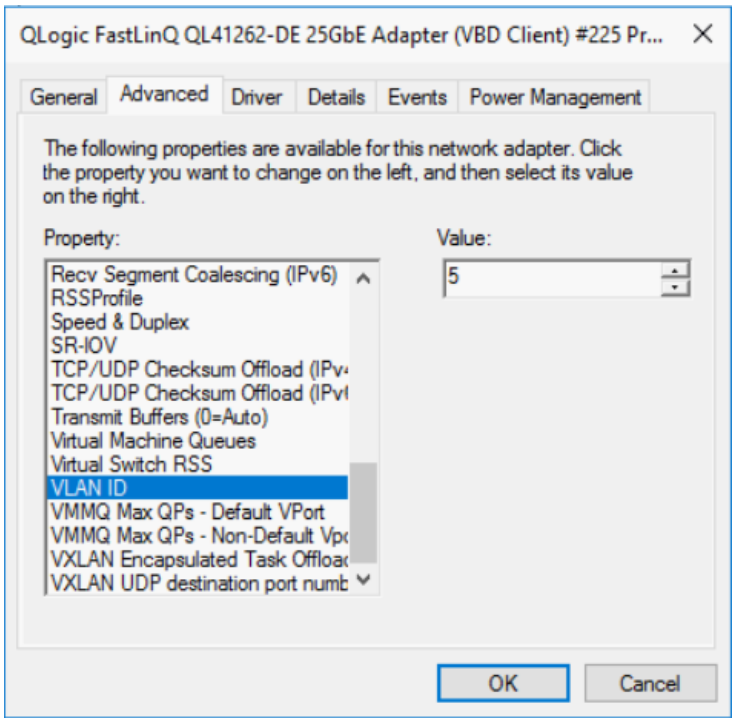


图 15-12. 高级属性：设置 VLAN ID

7. 要配置交换机，请发出以下 Windows PowerShell 命令：

```
PS C:\Users\Administrators> Get-NetAdapterQoS

Name                          : Ethernet 5
Enabled                       : True
Capabilities                  :                           Hardware     Current
                                                          --------     -------
                                MacSecBypass              : NotSupported NotSupported
                                DcbxSupport               : CEE          CEE
                                NumTCs(Max/ETS/PFC)       : 4/4/4        4/4/4

OperationalTrafficClasses     : TC TSA    Bandwidth Priorities
                                -- ---    --------- ----------
                                 0 ETS    5%        0-4,6-7
                                 1 ETS    95%       5

OperationalFlowControl        : Priority 5 Enabled
```

```
OperationalClassifications : Protocol  Port/Type Priority
                             --------  --------- --------
                             NetDirect 445       5

RemoteTrafficClasses       : TC TSA    Bandwidth Priorities
                             -- ---    --------- ----------
                              0 ETS    5%        0-4,6-7
                              1 ETS    95%       5

RemoteFlowControl          : Priority 5 Enabled
RemoteClassifications      : Protocol  Port/Type Priority
                             --------  --------- --------
                             NetDirect 445       5
```

注

上例为适配器端口连接到 Arista 7060X 交换机时采用。在此示例中，交换机 PFC 启用为优先级 5。RoCE App TLV 已定义。两个流量类定义为 TC0 和 TC1，其中 TC1 定义用于 RoCE。**DCBX Protocol**（DCBX 协议）模式设置为 **CEE**。有关 Arista 交换机配置，请参阅第 128 页上 “ 准备以太网交换机 ”。适配器处于 **Willing**（愿意）模式时，它会接受远程配置并将其显示为 **Operational Parameters**（操作参数）。

配置 VMMQ

虚拟机多队列 (VMMQ) 配置信息包括：

- 在适配器上启用 VMMQ
- 创建带或不带 SR-IOV 的虚拟机交换机
- 在虚拟机交换机上启用 VMMQ
- 获取虚拟机交换机功能
- 创建 VM 并在 VM 中的 VMNetworkAdapters 上启用 VM
- 在管理 NIC 上启用和禁用 VMMQ
- 监测流量统计信息

在适配器上启用 VMMQ

要在适配器上启用 VMMQ：

1. 打开微型端口属性，然后单击 **Advanced**（高级）选项卡。
2. 在适配器属性的 Advanced（高级）页面（图 15-13）的 **Property**（属性）下，选择 **Virtual Switch RSS**（虚拟交换机 RSS），然后将值设置为 **Enabled**（已启用）。
3. 单击 **OK**（确定）。

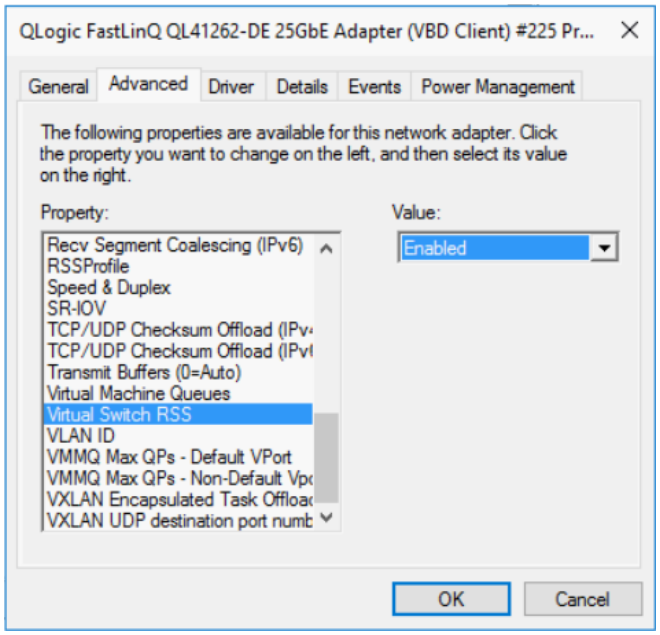


图 15-13. 高级属性：启用虚拟交换机 RSS

创建带或不带 SR-IOV 的虚拟机交换机

要创建带或不带 SR-IOV 的虚拟机交换机：

1. 启动 Hyper-V 管理器。
2. 选择 **Virtual Switch Manager**（虚拟交换机管理器）（请参阅图 15-14）。
3. 在 **Name**（名称）框中，键入虚拟交换机的名称。
4. 在 **Connection type**（连接类型）下：
   a. 单击 **External network**（外部网络）。
   b. 选择 **Allow management operating system to share this network adapter**（允许管理操作系统以共享此网络适配器）复选框。

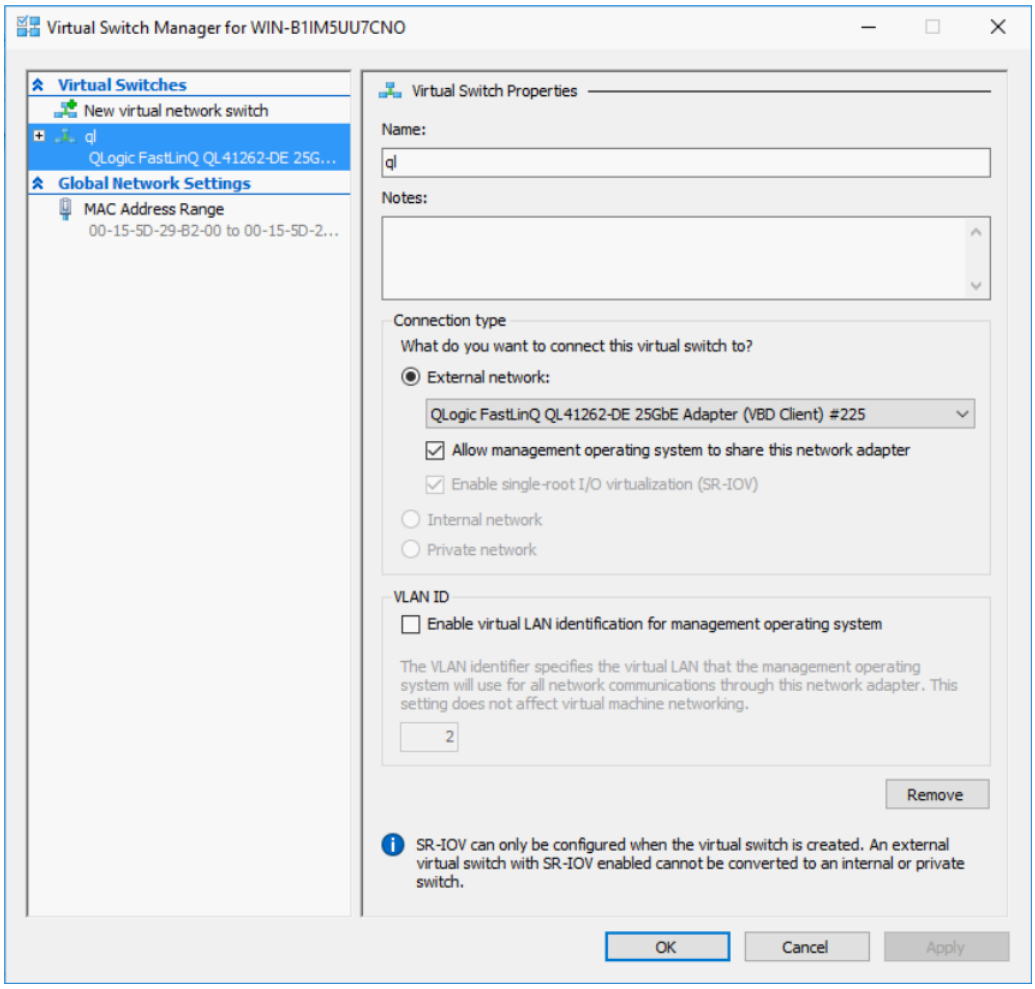


图 15-14. 虚拟交换机管理器

5. 单击 **OK**（确定）。

在虚拟机交换机上启用 VMMQ

要在虚拟机交换机上启用 VMMQ：

- 发出以下 Windows PowerShell 命令：

```
PS C:\Users\Administrators> Set-VMSwitch -name q1
-defaultqueuevmmqenabled $true -defaultqueuevmmqqueuepairs 4
```

获取虚拟机交换机功能

要获取虚拟机交换机功能：

- 发出以下 Windows PowerShell 命令：

  ```
  PS C:\Users\Administrator> Get-VMSwitch -Name ql | fl
  ```

  图 15-15 显示示例输出。

```
PS C:\Users\Administrator> Get-VMSwitch -Name ql | fl


Name                                : ql
Id                                  : 4dff5da3-f8bc-4146-a809-e1ddc6a04f7a
Notes                               :
Extensions                          : {Microsoft Windows Filtering Platform, Microsoft Azure VFP Switch Extension,
                                      Microsoft NDIS Capture}
BandwidthReservationMode            : None
PacketDirectEnabled                 : False
EmbeddedTeamingEnabled              : False
IovEnabled                          : True
SwitchType                          : External
AllowManagementOS                   : True
NetAdapterInterfaceDescription      : QLogic FastLinQ QL41262-DE 25GbE Adapter (VBD Client) #225
NetAdapterInterfaceDescriptions     : {QLogic FastLinQ QL41262-DE 25GbE Adapter (VBD Client) #225}
IovSupport                          : True
IovSupportReasons                   :
AvailableIPSecSA                    : 0
NumberIPSecSAAllocated              : 0
AvailableVMQueues                   : 103
NumberVmqAllocated                  : 1
IovQueuePairCount                   : 127
IovQueuePairsInUse                  : 2
IovVirtualFunctionCount             : 96
IovVirtualFunctionsInUse            : 0
PacketDirectInUse                   : False
DefaultQueueVrssEnabledRequested    : True
DefaultQueueVrssEnabled             : True
DefaultQueueVmmqEnabledRequested    : False
DefaultQueueVmmqEnabled             : False
DefaultQueueVmmqQueuePairsRequested : 16
DefaultQueueVmmqQueuePairs          : 16
BandwidthPercentage                 : 0
DefaultFlowMinimumBandwidthAbsolute : 0
DefaultFlowMinimumBandwidthWeight   : 0
CimSession                          : CimSession: .
ComputerName                        : WIN-B1IM5UU7CNO
IsDeleted                           : False
```

图 15-15. Windows PowerShell 命令：Get-VMSwitch

创建 VM 并在 VM 中的 VMNetworkAdapters 上启用 VM

要创建虚拟机 (VM) 并在 VM 的 VMNetworkadapter 中启用 VMMQ：

1. 创建 VM。
2. 将 VMNetworkadapter 添加到 VM。
3. 将虚拟交换机分配给 VMNetworkadapter。
4. 要在 VM 上启用 VMMQ，请发出以下 Windows PowerShell 命令：

   ```
   PS C:\Users\Administrators> set-vmnetworkadapter -vmname vm1
   -VMNetworkAdapterName "network adapter" -vmmqenabled $true
   -vmmqqueuepairs 4
   ```

在管理 NIC 上启用和禁用 VMMQ

要在管理 NIC 上启用或禁用 VMMQ：

- 要在管理 NIC 上启用 VMMQ，请发出以下命令：

  ```
  PS C:\Users\Administrator> Set-VMNetworkAdapter –ManagementOS
  –vmmqEnabled $true
  ```

- 要在管理 NIC 上禁用 VMMQ，请发出以下命令：

  ```
  PS C:\Users\Administrator> Set-VMNetworkAdapter –ManagementOS
  –vmmqEnabled $false
  ```

  VMMQ 也适用于多播优先打开最短路径 (MOSPF)。

监测流量统计信息

要监测虚拟机中的虚拟功能流量，请发出以下 Windows PowerShell 命令：

```
PS C:\Users\Administrator> Get-NetAdapterStatistics | fl
```

注

Marvell 支持使用 Windows Server 2016 和 Windows Server 2019 新增的参数在虚拟端口上配置队列对最大数量。有关详细信息，请参阅 第 279 页上 “每个 VPort 的最大队列对数 (L2)”。

配置 Storage Spaces Direct

Windows Server 2016 引入了 Storage Spaces Direct，这样便可通过本地存储构建高度可用和可扩展的存储系统。有关更多信息，请参阅以下 Microsoft TechNet 链接：

https://technet.microsoft.com/en-us/windows-server-docs/storage/storage-spaces/storage-spaces-direct-windows-server-2016

配置硬件

图 15-16 显示硬件配置的示例。

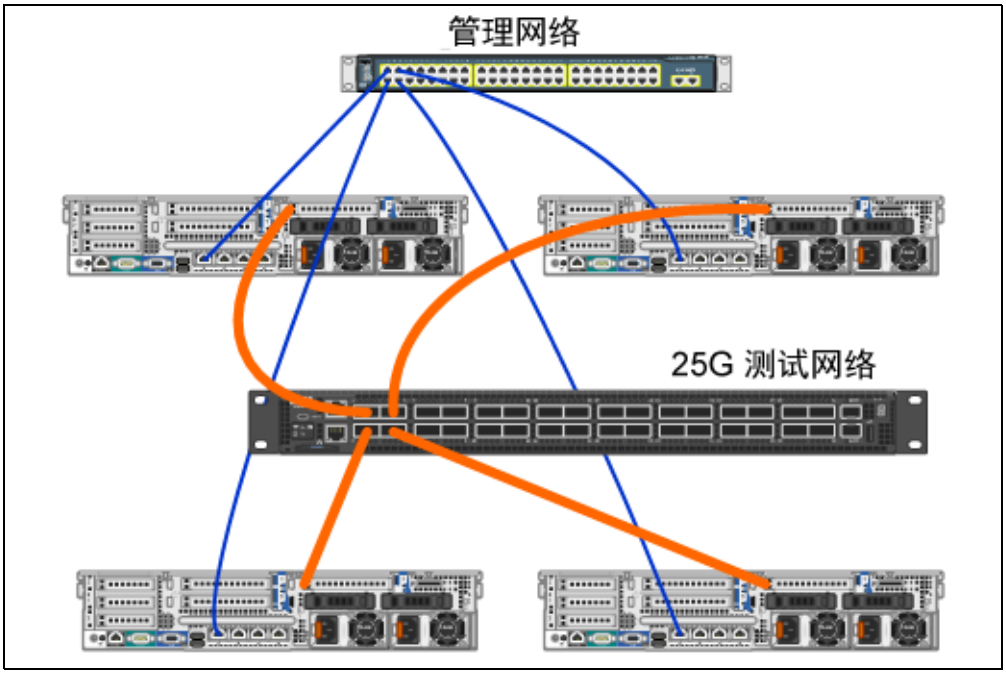


图 15-16. 示例硬件配置

注

本示例中使用的磁盘为 4 × 400G NVMe™ 和 12 × 200G SSD 磁盘。

部署超聚合系统

本节包括使用 Windows Server 2016 安装和配置超聚合系统组件的简介。部署超聚合系统的操作可以分为以下三个高级阶段：

- 部署操作系统
- 配置网络
- 配置 Storage Spaces Direct

部署操作系统

要部署操作系统：

1. 安装操作系统。
2. 安装 Windows 服务器角色 (Hyper-V)。

3. 安装以下功能：
   - ❑ 故障转移
   - ❑ 群集
   - ❑ 数据中心桥接 (DCB)
4. 将节点连接到域并添加域帐户。

配置网络

要部署 Storage Spaces Direct， Hyper-V 交换机必须部署为带有启用 RDMA 的主机虚拟 NIC。

注

以下步骤假设有四个 RDMA NIC 端口。

要在每个服务器上配置网络：

1. 请如下配置物理网络交换机：
   a. 将所有适配器 NIC 连接到交换机端口。

   注

   如果测试适配器有多个 NIC 端口，您必须将两个端口连接到同一交换机。

   b. 启用交换机端口并确保：
      - ■ 交换机端口支持交换机独立的组队模式。
      - ■ 交换机端口是多个 vLAN 网络的一部分。

      示例 Dell 交换机配置：

```
no ip address
mtu 9416
portmode hybrid
switchport
dcb-map roce_S2D
protocol lldp
dcbx version cee
no shutdown
```

2. 启用 **Network Quality of Service**（网络服务质量）。

注

网络服务质量用于确保软件定义的存储系统有足够的带宽在节点之间通信，以确保弹性和性能。要在适配器上配置 QoS，请参阅第 256 页上 “为 RoCE 配置 QoS”。

3. 请如下创建带 Switch Embedded Teaming (SET) 和 RDMA 虚拟 NIC 的 Hyper-V 虚拟交换机：

   a. 要标识网络适配器，请发出以下命令：

   ```
   Get-NetAdapter | FT
   Name,InterfaceDescription,Status,LinkSpeed
   ```

   b. 要创建连接到所有物理网络适配器的虚拟交换机，然后启用 SET，请发出以下命令：

   ```
   New-VMSwitch -Name SETswitch -NetAdapterName
   "<port1>","<port2>","<port3>","<port4>"
   -EnableEmbeddedTeaming $true
   ```

   c. 要将主机虚拟 NIC 添加到虚拟交换机，请发出以下命令：

   ```
   Add-VMNetworkAdapter –SwitchName SETswitch –Name SMB_1
   –managementOS
   Add-VMNetworkAdapter –SwitchName SETswitch –Name SMB_2
   –managementOS
   ```

   注

   上述命令从您刚刚为要使用的管理操作系统配置的虚拟交换机配置虚拟 NIC。

   d. 要配置主机虚拟 NIC 以使用 vLAN，请发出以下命令：

   ```
   Set-VMNetworkAdapterVlan -VMNetworkAdapterName "SMB_1"
   -VlanId 5 -Access -ManagementOS
   Set-VMNetworkAdapterVlan -VMNetworkAdapterName "SMB_2"
   -VlanId 5 -Access -ManagementOS
   ```

   注

   这些命令可位于相同或不同的 vLAN 上。

e. 要验证 vLAN ID 是否已设置，请发出以下命令：

```
Get-VMNetworkAdapterVlan -ManagementOS
```

f. 要禁用和启用每个主机虚拟 NIC 适配器以使 vLAN 处于活动状态，请发出以下命令：

```
Disable-NetAdapter "vEthernet (SMB_1)"
Enable-NetAdapter "vEthernet (SMB_1)"
Disable-NetAdapter "vEthernet (SMB_2)"
Enable-NetAdapter "vEthernet (SMB_2)"
```

g. 要在主机虚拟 NIC 适配器上启用 RDMA，请发出以下命令：

```
Enable-NetAdapterRdma "SMB1","SMB2"
```

h. 要验证 RDMA 功能，请发出以下命令：

```
Get-SmbClientNetworkInterface | where RdmaCapable -EQ
$true
```

配置 Storage Spaces Direct

在 Windows Server 2016 中配置 Storage Spaces Direct 包括以下步骤：

- 步骤 1. 运行群集验证工具
- 步骤 2. 创建群集
- 步骤 3. 配置群集见证
- 步骤 4. 清理用于 Storage Spaces Direct 的磁盘
- 步骤 5. 启用 Storage Spaces Direct
- 步骤 6. 创建虚拟磁盘
- 步骤 7. 创建或部署虚拟机

步骤 1. 运行群集验证工具

运行群集验证工具以确保服务器节点正确配置，从而使用 Storage Spaces Direct 创建群集。

要验证用作 Storage Spaces Direct 群集的一组服务器，请发出以下 Windows PowerShell 命令：

```
Test-Cluster -Node <MachineName1, MachineName2, MachineName3,
MachineName4> -Include "Storage Spaces Direct", Inventory,
Network, "System Configuration"
```

步骤 2. 创建群集

创建步骤 1. 运行群集验证工具中具有四个节点的群集 （群集创建已验证）。

要创建群集，请发出以下 Windows PowerShell 命令。

```
New-Cluster -Name <ClusterName> -Node <MachineName1, MachineName2,
MachineName3, MachineName4> -NoStorage
```

`-NoStorage` 参数是必需的。如果未包括该参数，磁盘将自动添加到群集，您必须移除磁盘，然后才能启用 Storage Spaces Direct。否则，磁盘将不包括在 Storage Spaces Direct 存储池中。

步骤 3. 配置群集见证

您应该配置群集的见证，以使该四节点系统能够承受两个节点发生故障或脱机。对于这些系统，您可配置文件共享见证或云见证。

有关详细信息，请访问：

https://docs.microsoft.com/en-us/windows-server/failover-clustering/manage-cluster-quorum

步骤 4. 清理用于 Storage Spaces Direct 的磁盘

旨在用于 Storage Spaces Direct 的磁盘必须为空，且不带分区或其他数据。如果磁盘有分区或其他数据，该磁盘将不会包括在 Storage Spaces Direct 系统中。

可在 Windows PowerShell 脚本 (`.PS1`) 文件中放入以下 Windows PowerShell 命令，并从管理系统打开的 Windows PowerShell （或 Windows PowerShell ISE）控制台以管理员权限执行。

注

运行此脚本可帮助识别每个节点上可用于 Storage Spaces Direct 的磁盘，并从这些磁盘移除所有数据和分区。

```
icm (Get-Cluster -Name HCNanoUSClu3 | Get-ClusterNode) {
Update-StorageProviderCache

Get-StoragePool |? IsPrimordial -eq $false | Set-StoragePool
-IsReadOnly:$false -ErrorAction SilentlyContinue

Get-StoragePool |? IsPrimordial -eq $false | Get-VirtualDisk |
Remove-VirtualDisk -Confirm:$false -ErrorAction SilentlyContinue

Get-StoragePool |? IsPrimordial -eq $false | Remove-StoragePool
-Confirm:$false -ErrorAction SilentlyContinue
```

```
Get-PhysicalDisk | Reset-PhysicalDisk -ErrorAction
SilentlyContinue

Get-Disk |? Number -ne $null |? IsBoot -ne $true |? IsSystem -ne
$true |? PartitionStyle -ne RAW |% {
$_ | Set-Disk -isoffline:$false
$_ | Set-Disk -isreadonly:$false
$_ | Clear-Disk -RemoveData -RemoveOEM -Confirm:$false
$_ | Set-Disk -isreadonly:$true
$_ | Set-Disk -isoffline:$true
}
Get-Disk |? Number -ne $null |? IsBoot -ne $true |? IsSystem -ne
$true |? PartitionStyle -eq RAW | Group -NoElement -Property
FriendlyName

} | Sort -Property PsComputerName,Count
```

步骤 5. 启用 Storage Spaces Direct

创建群集后，使用 `Enable-ClusterStorageSpacesDirect` Windows PowerShell cmdlet。该 cmdlet 会将存储系统置于 Storage Spaces Direct 模式并自动执行以下操作：

- 创建一个名如 *S2D on Cluster1* 的大型池。
- 配置 Storage Spaces Direct 高速缓存。如果有多种介质类型可供 Storage Spaces Direct 使用，它会配置最有效的类型作为高速缓存设备 （在大多数情况下，读取和写入）。
- 创建两个层作为默认层：**Capacity** （容量）和 **Performance** （性能）。该 cmdlet 会分析设备并以混合的设备类型和弹性配置每个层。

步骤 6. 创建虚拟磁盘

如果 Storage Spaces Direct 已启用，它会使用所有磁盘创建一个池。它还会命名该池 （例如 *S2D on Cluster1*），并在名称中指定群集名称。

以下 Windows PowerShell 命令在存储池上创建具有镜像和奇偶校验弹性的虚拟磁盘：

```
New-Volume -StoragePoolFriendlyName "S2D*" -FriendlyName
<VirtualDiskName> -FileSystem CSVFS_ReFS -StorageTierfriendlyNames
Capacity,Performance -StorageTierSizes <Size of capacity tier in
size units, example: 800GB>, <Size of Performance tier in size
units, example: 80GB> -CimSession <ClusterName>
```

步骤 7. 创建或部署虚拟机

您可以在超聚合 *S2D* 群集的节点上部署虚拟机。将虚拟机的文件存储在系统的 Cluster Shared Volume (CSV) 名称空间 （例如，`c:\ClusterStorage\Volume1`）中，与故障转移群集上的群集虚拟机类似。

16 Windows Server 2019

本章提供有关 Windows Server 2019 的以下信息：

Hyper-V 的 RSSv2

在 Windows Server 2019 中， Microsoft 增加了对包含 Hyper-V （每个 vPort 的 RSSv2）的接收方缩放第 2 版 (RSSv2) 的支持。

RSSv2 说明

与 RSSv1 相比， RSSv2 可减少 CPU 负载测量与间接表更新之间的时间。此功能可防止高流量情况下速度变慢。 RSSv2 可以跨多个处理器动态传播接收队列，响应速度远远高于 RSSv1。有关更多信息，请访问以下网页：

https://docs.microsoft.com/en-us/windows-hardware/drivers/network/receive-side-scaling-version-2-rssv2-

当 **Virtual Switch RSS** （虚拟交换机 RSS）选项也已启用时， Windows Server 2019 驱动程序中默认支持 RSSv2。此选项默认启用，并且 NIC 绑定到 Hyper-V 或 vSwitch。

已知事件日志错误

在常见操作下， RSSv2 的动态算法可能会启动与驱动程序不兼容的间接表更新，并返回适当的状态代码。在这种情况下，会出现事件日志错误，即使不存在功能操作问题也一样。图 16-1 显示一个示例。

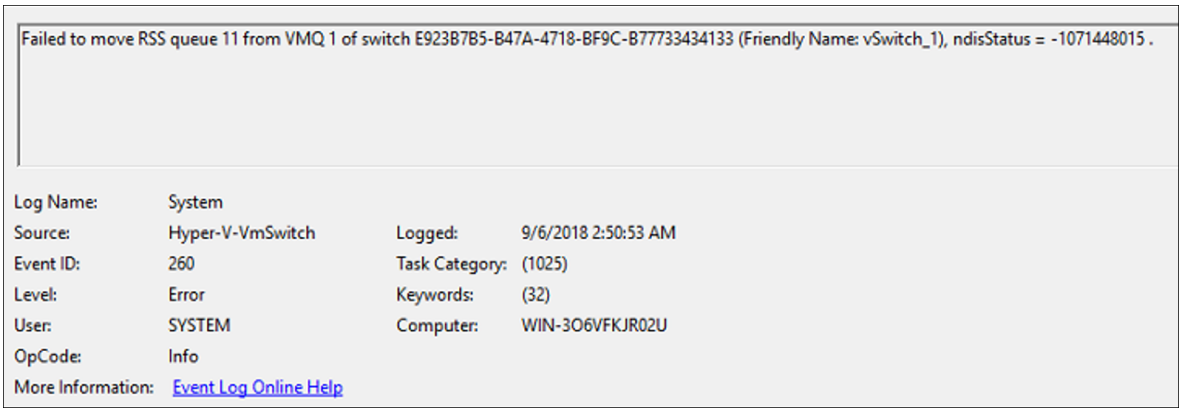


图 16-1. RSSv2 事件日志错误

Windows Server 2019 行为

Windows Server 2019 引入了会影响适配器配置的以下新行为。

VMMQ 默认启用

在 Windows Server 2019 的内建驱动程序中， **Virtual Switch RSS** （虚拟交换机 RSS） (VMMQ) 选项默认在 NIC 属性中启用。此外， Microsoft 更改了 **Virtual NICs** （虚拟 NIC）选项的默认行为，让 VMMQ 以 16 个队列对启用。此行为更改会影响可用资源的数量。

例如，假设 NIC 支持 32 个 VMQ 和 64 个队列对。在 Windows Server 2016 中，当您增加 32 个虚拟 NIC (VNIC) 时，它们将有 VMQ 加速。但在 Windows Server 2019 中，您将获得 4 个 （每个有 16 个队列对）有 VMMQ 加速的 VNIC，以及 30 个没有加速的 VNIC。

由于此功能的作用， Marvell 引入了新的用户属性 **Max Queue Pairs (L2) Per VPort**。有关更多详情，请参阅 新适配器属性。

内建驱动程序 Network Direct (RDMA) 默认禁用

在 Windows Server 2019 的内建驱动程序中， **Network Direct** (RDMA) 选项默认在 NIC 属性中禁用。但在将驱动程序升级到开箱即用驱动程序时， **Network Direct** 默认启用。

新适配器属性

以下章节介绍 Windows Server 2019 中新增的用户可配置属性：

- 每个 VPort 的最大队列对数 (L2)
- Network Direct 技术
- 虚拟化资源
- VMQ 和 VMMQ 默认加速
- 单一 VPort 池

每个 VPort 的最大队列对数 (L2)

如 VMMQ 默认启用中所述， Windows 2019 （和 Windows 2016）新增了用户可配置参数 **Max Queue Pairs (L2) per VPort** （每个 VPort 的最大队列对数 (L2)）。此参数允许定义可分配到以下项目的最大队列对数量，以增强对资源分发的控制：

- VPort- 默认 VPort
- PF 非默认 VPort (VMQ/VMMQ)
- SR-IOV 非默认 VPort (VF)[1]

Max Queue Pairs (L2) per VPort （每个 VPort 的最大队列对数 (L2)）参数的默认值设为 **Auto** （自动），为以下值之一：

- 默认 vPort 的最大队列对数 = 8
- 非默认 vPort 的最大队列对数 = 4

如果选择小于 8 的值，则：

- 默认 vPort 的最大队列对数 = 8
- 非默认 vPort 的最大队列对数 = 值

如果选择大于 8 的值，则：

- 默认 vPort 的最大队列对数 = 值
- 非默认 vPort 的最大队列对数 = 值

Network Direct 技术

Marvell 支持新的 **Network Direct 技术** 参数，该参数可让您选择符合以下 Microsoft 规格的底层 RDMA 技术：

https://docs.microsoft.com/en-us/windows-hardware/drivers/network/inf-requirements-for-ndkpi

此选项替换 **RDMA Mode** （RDMA 模式）参数。

[1] 此参数也适用于 Windows Server 2016。

虚拟化资源

表 16-1 列出了 Windows 2019 中用于 Dell 41xxx 系列适配器的最大虚拟化资源数。

表 16-1. 用于 Dell 41xxx 系列适配器的 Windows 2019 虚拟化资源

双端口 NIC- 仅单一功能非 -CNA	数量
最大 VMQ 数	102
最大 VF 数	80
最大 QP 数	112
四端口 NIC- 仅单一功能非 -CNA	**数量**
最大 VMQ 数	47
最大 VF 数	32
最大 QP 数	48

VMQ 和 VMMQ 默认加速

表 16-2 列出了 Windows Server 2019 中用于 Dell 41xxx 系列适配器的 VMQ 和 VMMQ 默认值及其他值。

表 16-2. Windows 2019 VMQ 和 VMMQ 加速

双端口 NIC- 仅单一功能非 -CNA	默认值	其他可能值				
每个 VPort 的最大队列对数 (L2) [a]	自动	1	2	4	8	16
最大 VMQ 数	26	103	52	26	13	6
默认 Vport 队列对数	8	8	8	8	8	16
PF 非默认 Vport 队列对数	4	1	2	4	8	16
四端口 NIC- 仅单一功能非 -CNA	**默认值**	**其他可能值**				
每个 VPort 的最大队列对数 (L2)[a]	自动	1	2	4	8	16
最大 VMQ 数	10	40	20	10	5	2
默认 Vport 队列对数	8	8	8	8	8	16
PF 非默认 Vport 队列对数	4	1	2	4	8	16

[a] **Max Queue Pairs (L2) VPort** （VPort 的最大队列对数 (L2)）是 NIC 高级属性的可配置参数。

单一 VPort 池

41*xxx* 系列适配器 支持 **Single VPort Pool** （单一虚拟端口池）参数，可让系统管理员将任何可用的 IOVQueuePair 分配到默认 VPort、 PF 非默认 VPort 或 VF 非默认 VPort。要分配值，请发出以下 Windows PowerShell 命令：

- 默认 VPort：

```
Set-VMSwitch -Name <vswitch name> -DefaultQueueVmmqEnabled:1
-DefaultQueueVmmqQueuePairs:<number>
```

注

Marvell 不建议禁用 VMMQ 或减少默认 VPort 的队列对数量，因为可能影响系统性能。

- PF 非默认 VPort：
  - 对于主机：

    ```
    Set-VMNetworkAdapter -ManagementOS -VmmqEnabled:1
    -VmmqQueuePairs:<number>
    ```

  - 对于 VM：

    ```
    Set-VMNetworkAdapter -VMName <vm name> -VmmqEnabled:1
    -VmmqQueuePairs:<number>
    ```

- VF 非默认 VPort：

  ```
  Set-VMNetworkAdapter -VMName <vm name> -IovWeight:100
  -IovQueuePairsRequested:<number>
  ```

  注

  为 VF 分配的默认 QP 数 (**IovQueuePairsRequested**) 仍为 1。

要将大量队列对应用到任何 vPort：

- 队列对数量必须小于或等于系统上 CPU 核心的总数。
- 队列对数量必须小于或等于 **Max Queue Pairs (L2) Per VPort**（每个 VPort 的最大队列对数 (L2)）的值。有关更多信息，请参阅 每个 VPort 的最大队列对数 (L2)。

17 故障排除

本章提供了以下故障排除信息：

故障排除核查表

小心

在打开服务器机箱以添加或拆卸适配器之前，请先查阅第 5 页上 “ 安全预防措施 ”。

以下核查表提供了一些建议的操作，以解决在系统中安装或运行 41*xxx* 系列适配器时可能遇到的问题。

- 检查所有电缆和连接。验证网络适配器和交换机上的电缆正确连接。
- 对照第 6 页上 “ 安装适配器 ” 验证适配器安装。确保适配器正确插入插槽中。检查是否有特定的硬件问题，如插卡组件或 PCI 边缘连接器明显损坏。
- 验证配置设置，如果与其他设备冲突，则进行更改。
- 验证服务器在使用最新的 BIOS。
- 尝试将适配器插入另一插槽。如果在新位置没有问题，则系统中的原插槽可能有缺陷。
- 用已知工作正常的适配器替换故障适配器。如果第二个适配器在第一个适配器无法运行的插槽中可运行，则原适配器可能有缺陷。
- 将适配器安装在另一台运行正常的系统中，然后再次运行测试。如果适配器在新系统中通过测试，则原系统可能有缺陷。

- 从该系统卸下所有其他适配器，然后再次运行测试。如果适配器通过测试，则其他适配器可能导致了争用。

验证是否已加载最新驱动程序

确保已为您的 Windows、 Linux 或 VMware 系统加载最新驱动程序。

验证 Windows 中的驱动程序

请参阅设备管理器，查看有关适配器、链路状态和网络连接的重要信息。

验证 Linux 中的驱动程序

要验证 qed.ko 驱动程序是否已正确加载，请发出以下命令：

```
# lsmod | grep -i <module name>
```

如果驱动程序已加载，此命令的输出将显示驱动程序大小 （以字节为单位）。以下示例显示为 qed 模块加载的驱动程序：

```
# lsmod | grep -i qed
qed                   199238  1
qede                 1417947  0
```

如果加载新的驱动程序后重新引导，可以发出以下命令验证当前加载的驱动程序版本是否正确：

```
modinfo qede
```

也可以发出以下命令：

```
[root@test1]# ethtool -i eth2
driver: qede
version: 8.4.7.0
firmware-version: mfw 8.4.7.0 storm 8.4.7.0
bus-info: 0000:04:00.2
```

如果已加载新驱动程序，但尚未重新引导，则 `modinfo` 命令不会显示更新的驱动程序信息。相反地，发出以下 `dmesg` 命令可查看日志。在此示例中，最后一个条目用于标识将在重新引导时激活的驱动程序。

```
# dmesg | grep -i "Cavium" | grep -i "qede"

[   10.097526] QLogic FastLinQ 4xxxx Ethernet Driver qede x.x.x.x
[   23.093526] QLogic FastLinQ 4xxxx Ethernet Driver qede x.x.x.x
[   34.975396] QLogic FastLinQ 4xxxx Ethernet Driver qede x.x.x.x
[   34.975896] QLogic FastLinQ 4xxxx Ethernet Driver qede x.x.x.x
[ 3334.975896] QLogic FastLinQ 4xxxx Ethernet Driver qede x.x.x.x
```

验证 VMware 中的驱动程序

要验证 VMware ESXi 驱动程序已加载，请发出以下命令：

```
# esxcli software vib list
```

测试网络连接

本节提供在 Windows 和 Linux 环境中测试网络连接的步骤。

注

在使用强制链路速度时，验证适配器和交换机均被强制为同一速度。

测试 Windows 的网络连接

使用 `ping` 命令测试网络连接。

要确定网络连接是否工作：

1. 单击 **Start**（开始），然后单击 **Run**（运行）。
2. 在 **Open**（打开）框中，键入 **cmd**，然后单击 **OK**（确定）。
3. 要查看被测试的网络连接，发出以下命令：

   ```
   ipconfig /all
   ```

4. 发出以下命令，然后按 ENTER。

   ```
   ping <ip_address>
   ```

显示的 ping 统计信息指示网络连接是否在工作。

测试 Linux 的网络连接

要验证以太网接口正常工作并运行：

1. 要检查以太网接口的状态，请发出 `ifconfig` 命令。
2. 要检查以太网接口的统计信息，请发出 `netstat -i` 命令。

要验证连接是否已建立：

1. 对网络上的 IP 主机执行 Ping 操作。从命令行，发出以下命令：

   ```
   ping <ip_address>
   ```

2. 按 ENTER 键。

显示的 ping 统计信息指示网络连接是否在工作。

使用操作系统 GUI 工具或 ethtool 命令 `ethtool -s ethX speed SSSS` 可以将适配器链路速度强制为 10 Gbps 或 25 Gbps。

使用 Hyper-V 的 Microsoft 虚拟化

Microsoft 虚拟化是适用于 Windows Server 2012 R2 的虚拟机监控程序虚拟化系统。有关 Hyper-V 的更多信息，请访问：

https://technet.microsoft.com/en-us/library/Dn282278.aspx

Linux 特定问题

问题： 汇编驱动程序源码时出错。

解决方案： Linux 分发版的有些安装没有默认安装开发工具和内核源。汇编驱动程序源码之前，确保所使用的 Linux 分发版的开发工具已安装。

其他问题

问题： 41*xxx* 系列适配器已关闭并出现错误消息，提示适配器上的风扇发生故障。

解决方案： 可人为关闭 41*xxx* 系列适配器，以防止永久损坏。联系 Marvell 技术支持以获取帮助。

问题： 在安装有 iSCSI 驱动程序 (qedil) 的 ESXi 环境中，有时 VI 客户端无法访问主机。这是由于 hostd 守护程序终止，影响了与 VI 客户端的连接性。

解决方案： 联系 VMware 技术支持

收集调试数据

使用 表 17-1 中的命令收集调试数据。

表 17-1. 收集调试数据命令

调试数据	说明
demesg-T	内核日志
ethtool-d	寄存器转储
sys_info.sh	系统信息。此项在驱动程序包中提供。

A 适配器 LED

表 A-1 列出适配器端口链路和活动状态的 LED 指示灯。

表 A-1. 适配器端口链路和活动 LED

端口 LED	LED 外观	网络状态
链路 LED	不亮	无链接 （电缆断开或端口关闭）
	持续绿色发亮	以最高支持的链路速度链接
	持续黄褐色发亮	以最低支持的链路速度链接
活动 LED	不亮	无端口活动
	闪烁	端口活动

B 电缆和光学模块

本附录提供有关支持的电缆和光学模块的以下信息：

支持的规格

41*xxx* 系列适配器支持符合 SFF8024 的各种电缆和光学模块。具体的外形规格符合性如下：

- SFP：
  - SFF8472 （用于内存映射）
  - SFF8419 或 SFF8431 （低速信号和电源）
- 光学模块电气输入 / 输出、有源铜缆 (ACC) 和有源光缆 (AOC)：
  - 10G—SFF8431 限制接口
  - 25G—IEEE 802.3by Annex 109B (25GAUI) （不支持 RS-FEC）

测试的电缆和光学模块

Marvell 不保证满足符合性要求的每个电缆或光学模块均可与 41*xxx* 系列适配器一起使用。 Marvell 已测试表 B-1 中列出的部件并提供此列表方便您使用。

表 B-1. 测试的电缆和光学模块

速度 / 形状因素	制造商	部件号	类型	电缆长度[a]	标准尺寸
电缆					
10G DAC[b]	Brocade®	1539W	SFP+10G-to-SFP+10G	1	26
		V239T	SFP+10G-to-SFP+10G	3	26
		48V40	SFP+10G-to-SFP+10G	5	26
	Cisco	H606N	SFP+10G-to-SFP+10G	1	26
		K591N	SFP+10G-to-SFP+10G	3	26
		G849N	SFP+10G-to-SFP+10G	5	26
	Dell	V250M	SFP+10G-to-SFP+10G	1	26
		53HVN	SFP+10G-to-SFP+10G	3	26
		358VV	SFP+10G-to-SFP+10G	5	26
		407-BBBK	SFP+10G-to-SFP+10G	1	30
		407-BBBI	SFP+10G-to-SFP+10G	3	26
		407-BBBP	SFP+10G-to-SFP+10G	5	26
25G DAC	Amphenol®	NDCCGF0001	SFP28-25G-to-SFP28-25G	1	30
		NDCCGF0003	SFP28-25G-to-SFP28-25G	3	30
		NDCCGJ0003	SFP28-25G-to-SFP28-25G	3	26
		NDCCGJ0005	SFP28-25G-to-SFP28-25G	5	26
	Dell	2JVDD	SFP28-25G-to-SFP28-25G	1	26
		D0R73	SFP28-25G-to-SFP28-25G	2	26
		OVXFJY	SFP28-25G-to-SFP28-25G	3	26
		9X8JP	SFP28-25G-to-SFP28-25G	5	26

表 B-1. 测试的电缆和光学模块 (续)

速度 / 形状因素	制造商	部件号	类型	电缆长度[a]	标准尺寸
40G 铜质 QSFP 分配器 (4×10G)	Dell	TCPM2	QSFP+40G-to-4xSFP+10G	1	30
		27GG5	QSFP+40G-to-4xSFP+10G	3	30
		P8T4W	QSFP+40G-to-4xSFP+10G	5	26
1G 铜质 RJ45 收发器	Dell	8T47V	SFP+ to 1G RJ	1G RJ45	N/A
		XK1M7	SFP+ to 1G RJ	1G RJ45	N/A
		XTY28	SFP+ to 1G RJ	1G RJ45	N/A
10G 铜质 RJ45 收发器	Dell	PGYJT	SFP+ to 10G RJ	10G RJ45	N/A
40G DAC 分配器 (4×10G)	Dell	470-AAVO	QSFP+40G-to-4xSFP+10G	1	26
		470-AAXG	QSFP+40G-to-4xSFP+10G	3	26
		470-BBBK	QSFP+40G-to-4xSFP+10G	5	26
100G DAC 分配器 (4×25G)	Amphenol	NDAQGJ-0001	QSFP28-100G-to-4xSFP28-25G	1	26
		NDAQGJ-0002	QSFP28-100G-to-4xSFP28-25G	2	30
		NDAQGF-0003	QSFP28-100G-to-4xSFP28-25G	3	30
		NDAQGJ-0005	QSFP28-100G-to-4xSFP28-25G	5	26
	Dell	026FN3 Rev A00	QSFP28-100G-to-4XSFP28-25G	1	26
		0YFNDD Rev A00	QSFP28-100G-to-4XSFP28-25G	2	26
		07R9N9 Rev A00	QSFP28-100G-to-4XSFP28-25G	3	26
	FCI	10130795-4050LF	QSFP28-100G-to-4XSFP28-25G	5	26

表 B-1. 测试的电缆和光学模块 (续)

速度 / 形状因素	制造商	部件号	类型	电缆长度[a]	标准尺寸
光学解决方案					
10G 光学收发器	Avago®	AFBR-703SMZ	SFP+ SR	N/A	N/A
		AFBR-701SDZ	SFP+ LR	N/A	N/A
	Dell	Y3KJN	SFP+ SR	1G/10G	N/A
		WTRD1	SFP+ SR	10G	N/A
		3G84K	SFP+ SR	10G	N/A
		RN84N	SFP+ SR	10G-LR	N/A
	Finisar®	FTLX8571D3BCL-QL	SFP+ SR	N/A	N/A
		FTLX1471D3BCL-QL	SFP+ LR	N/A	N/A
25G 光学收发器	Dell	P7D7R	SFP28 光学收发器 SR	25G SR	N/A
	Finisar	FTLF8536P4BCL	SFP28 光学收发器 SR	N/A	N/A
		FTLF8538P4BCL	SFP28 光学收发器 SR 非 FEC	N/A	N/A
10/25G 双速率收发器	Dell	M14MK	SFP28	N/A	N/A

表 B-1. 测试的电缆和光学模块 (续)

速度 / 形状因素	制造商	部件号	类型	电缆长度[a]	标准尺寸
10G AOC[c]	Dell	470-ABLV	SFP+ AOC	2	N/A
		470-ABLZ	SFP+ AOC	3	N/A
		470-ABLT	SFP+ AOC	5	N/A
		470-ABML	SFP+ AOC	7	N/A
		470-ABLU	SFP+ AOC	10	N/A
		470-ABMD	SFP+ AOC	15	N/A
		470-ABMJ	SFP+ AOC	20	N/A
		YJF03	SFP+ AOC	2	N/A
		P9GND	SFP+ AOC	3	N/A
		T1KCN	SFP+ AOC	5	N/A
		1DXKP	SFP+ AOC	7	N/A
		MT7R2	SFP+ AOC	10	N/A
		K0T7R	SFP+ AOC	15	N/A
		W5G04	SFP+ AOC	20	N/A
25G AOC	Dell	X5DH4	SFP28 AOC	20	N/A
	InnoLight®	TF-PY003-N00	SFP28 AOC	3	N/A
		TF-PY020-N00	SFP28 AOC	20	N/A

[a] 电缆长度以米为单位。

[b] DAC 是直接连接电缆。

[c] AOC 是有源光缆。

测试的交换机

表 B-2 列出已经通过与 41*xxx* 系列适配器互操作性测试的交换机。此列表基于产品发布时可用的交换机，且随着新交换机进入市场或停产，将随时间变化。

表 B-2. 进行互操作性测试的交换机

制造商	以太网交换机型号
Arista	7060X 7160
Cisco	Nexus 3132 Nexus 3232C Nexus 5548 Nexus 5596T Nexus 6000
Dell EMC	S6100 Z9100
HPE	FlexFabric 5950
Mellanox	SN2410 SN2700

C Dell Z9100 交换机配置

41*xxx* 系列适配器支持与 Dell Z9100 以太网交换机连接。但是，在标准化自动协商过程之前，必须将交换机明确地配置为按 25 Gbps 速率连接到适配器。

要配置 Dell Z9100 交换机端口以便按 25 Gbps 速率连接 41*xxx* 系列适配器：

1. 在您的管理工作站与交换机之间建立串行端口连接。
2. 打开命令行会话，然后如下登录到交换机：

```
Login: admin
Password: admin
```

3. 启用交换机端口的配置：

```
Dell> enable
Password: xxxxxx
Dell# config
```

4. 标识要配置的模块和端口。以下示例使用模块 1，端口 5：

```
Dell(conf)#stack-unit 1 port 5 ?
portmode                Set portmode for a module
Dell(conf)#stack-unit 1 port 5 portmode ?
dual                    Enable dual mode
quad                    Enable quad mode
single                  Enable single mode
Dell(conf)#stack-unit 1 port 5 portmode quad ?
speed                   Each port speed in quad mode
Dell(conf)#stack-unit 1 port 5 portmode quad speed ?
10G                     Quad port mode with 10G speed
25G                     Quad port mode with 25G speed
Dell(conf)#stack-unit 1 port 5 portmode quad speed 25G
```

有关更改适配器链路速度的信息，请参阅第 285 页上 “ 测试网络连接 ”。

5. 验证端口的运行速率是否为 25 Gbps：

```
Dell# Dell#show running-config | grep "port 5"
stack-unit 1 port 5 portmode quad speed 25G
```

6. 要在交换机端口 5 上禁用自动协商，请按照以下步骤操作：

   a. 标识交换机端口接口 （模块 1，端口 5，接口 1），并确认自动协商状态：

```
Dell(conf)#interface tw 1/5/1

Dell(conf-if-tf-1/5/1)#intf-type cr4 ?
autoneg                    Enable autoneg
```

   b. 禁用自动协商：

```
Dell(conf-if-tf-1/5/1)#no intf-type cr4 autoneg
```

   c. 验证是否已禁用自动协商。

```
Dell(conf-if-tf-1/5/1)#do show run interface tw 1/5/1
!
interface twentyFiveGigE 1/5/1
no ip address
mtu 9416
switchport
flowcontrol rx on tx on
no shutdown
no intf-type cr4 autoneg
```

有关配置 Dell Z9100 交换机的更多信息，请参阅 Dell 支持网站上的 *Dell Z9100 Switch Configuration Guide* （Dell Z9100 交换机配置指南），网址为：

support.dell.com

D 功能约束

本附录提供有关当前版本中实施的功能约束的信息。

未来的版本中可能会删除这些功能共存约束。到时，您应该可以使用功能组合，而无需超出启用功能通常所需的任何额外的配置步骤。

NPAR 模式下不支持同一端口上共存的 FCoE 和 iSCSI

在 NPAR 模式下，设备不支持在同一端口上配置 FCoE 卸载和 iSCSI 卸载。在 NPAR 模式下，第二个物理功能 (PF) 上 支持 FCoE 卸载，第三个 PF 上支持 iSCSI 下载。在单一 Ethernet PF DEFAULT Mode （以太网 PF 默认模式）下，设备支持在同一端口上配置 FCoE 卸载和 iSCSI 卸载。并非所有设备都支持 FCoE 卸载和 iSCSI 卸载。

使用 HII 或 Marvell 管理工具在一个端口上配置具有 iSCSI 或 FCoE 个性设置的 PF 后，禁止通过这些管理工具在另一个 PF 上配置存储协议。

由于存储个性设置默认已禁用，因此只有使用 HII 或 Marvell 管理工具配置的个性设置才能写入 NVRAM 配置中。移除此限制后，用户在 NPAR 模式下可在同一端口上配置其他 PF 用于存储。

不支持同一物理功能上共存的 RoCE 和 iWARP

不支持同一 PF 上的 RoCE 和 iWARP。 UEFI HII 和 Marvell 管理工具允许用户同时配置两者，但在这种情况下 RoCE 功能优先于 iWARP 功能，除非被 OS 中的驱动程序设置所覆盖。

仅在所选 PF 上支持 NIC 和 SAN 引导至库

目前只在物理端口的第一个以太网 PF 上支持以太网 （如软件 iSCSI 远程引导）和 PXE 引导。在 NPAR 模式配置中，第一个以太网 PF （不是另一个以太网 PF）支持以太网 （如软件 iSCSI 远程引导）和 PXE 引导。并非所有设备都支持 FCoE 卸载和 iSCSI 卸载。

- 当 **Virtualization** （虚拟化）或 **Multi-Function Mode** （多功能模式）设置为 **NPAR** 时，物理端口的第二个 PF 上支持 FCoE 卸载引导，物理端口的第三个 PF 上支持 iSCSI 卸载引导，而物理端口的第一个 PF 上同时支持以太网 （如软件 iSCSI）和 PXE 引导。
- iSCSI 和 FCoE 引导限于每个引导会话一个目标。
- 每个物理端口只允许一种引导模式。
- 仅 NPAR 模式才支持 iSCSI 卸载和 FCoE 卸载。

E 修订历史

文档修订历史	
修订版 A， 2017 年 4 月 28 日	
修订版 B， 2017 年 8 月 24 日	
修订版 C， 2017 年 10 月 1 日	
修订版 D， 2018 年 1 月 24 日	
修订版 E， 2018 年 3 月 15 日	
修订版 F， 2018 年 4 月 19 日	
修订版 G， 2018 年 5 月 22 日	
修订版 H， 2018 年 8 月 23 日	
修订版 J， 2019 年 1 月 23 日	
修订版 K， 2019 年 7 月 2 日	
修订版 L， 2019 年 7 月 3 日	
修订版 M， 2019 年 10 月 16 日	
更改	**受影响的章节**
将以下适配器添加到 Marvell 产品列表中：QL41164HFRJ-DE, QL41164HFRJ-DE, QL41164HFCU-DE, QL41232HMKR-DE, QL41262HMKR-DE, QL41232HFCU-DE, QL41232HLCU-DE, QL41132HFRJ-DE, QL41132HLRJ-DE, QL41132HQRJ-DE, QL41232HQCU-DE, QL41132HQCU-DE, QL41154HQRJ-DE, QL41154HQCU-DE	第 xvi 页上 “ 支持的产品 ”
增加了对 VMDirectPath I/O 的支持	第 1 页上 “ 功能 ”
在 表 2-2 中，更新了支持 Windows Server, RHEL, SLES, XenServer 的操作系统。	第 4 页上 “ 系统要求 ”

在第二段后面的项目符号中，添加了进一步描述 Dell iSCSI HW 和 SW 安装的文本。	第 66 页上 “iSCSI 预引导配置 ”
将该部分移动至更靠近其他相关部分。	第 69 页上 “ 配置存储目标 ”
在第一段中，将第一句更正为 “ **引导模式**选项列在 ***NIC* 配置下 ……**” 添加了设置 UEFI iSCSI HBA 的说明。	第 70 页上 “ 选择 iSCSI UEFI 引导协议 ”
删除了 “ 配置 iSCSI 启动参数 ” 和 “ 配置基本 BIOS 引导模式 ” 部分。	第 65 页上 “ 从 SAN 引导配置 ”
添加了 *Application Note, Enabling Storage Offloads on Dell and Marvell FastLinQ 41000 Series Adapters* 的参考资料。	第 36 页上 “FCoE 支持 ”, 第 36 页上 “iSCSI 支持 ”, 第 53 页上 “ 配置 FCoE 引导 ”, 第 55 页上 “ 配置 iSCSI 引导 ”, 第 65 页上 “ 从 SAN 引导配置 ”, 第 197 页上 “iSCSI 配置 ”, 第 210 页上 “FCoE 配置 ”
在 步骤 3 中，阐明了 RoCE v1 优先级值的使用方式。	第 52 页上 “ 配置数据中心桥接 ”
在本节末尾添加了一个注释，其中包含一个如何在激活 MPIO 配置和单路径的 Linux 中安装 iSCSI·BFS 的示例。	第 93 页上 “ 从 RHEL 7.5 及更高版本的 SAN 配置 iSCSI 引导 ”
更新了将适配器驱动程序滑流至窗口映像文件的步骤。	第 120 页上 “ 将适配器驱动程序注入 （滑流至） Windows 映像文件中 ”
删除了说明 “RoCE 无法在 SR-IOV 环境中的 VF 上工作 ” 的项目符号现已支持 VF RDMA。	第 127 页上 “ 计划 RoCE”
在 表 7-1 中， 移除了 RHEL 7.5 ；添加了 RHEL 7.7。对于 RHEL7.6，添加了对 OFED-4.17-1 GA 的支持。 移除了 SLES 12 SP3 ；添加了 SLES 12 SP4。 分离了 SLES 15 (SP0) 和 SLES 15 SP1 ；在 SP1 SLES 15 中，添加了对 OFED-4.17-1 GA 的支持。 CentOS 7.6：添加了对 OFED-4.17-1 GA 的支持。	第 126 页上 “ 支持的操作系统和 OFED”
添加了有关为 Windows 和 Linux 配置 RoCE 的 RDMA 的信息。	第 140 页上 “ 为 SR-IOV VF 设备 (VF RDMA) 配置 RoCE” 为 SR-IOV VF 设备 (VF RDMA) 配置 RoCE
在 步骤 1 中，分别将第二和第三个项目符号更新为目前支持 SLES 12 和 RHEL 的操作系统。	第 153 页上 “Linux 的 RoCE v2 配置 ”

将 步骤 4 部分 b 更改为 “ 将 **RDMA 协议支持**设置为 ***RoCE/iWARP*** 或 **iWARP**。 ”	第 174 页上 “ 为 iWARP 准备适配器 ”
删除对 附录 C 的引用；添加了配置信息。	第 130 页上 “ 为 RoCE 配置 Dell Z9100 以太网交换机 ”
支持内建 OFED 的操作系统更新列表。	第 185 页上 “ 准备工作 ”
删除了 “SLES 12 SP3 和 OFED 4.8*x* 上的 iWARP RDMA-Core 支持 ” 部分。	第 174 页上 “iWARP 配置 ”
在第三段后的项目符号列表中，更新了支持的操作系统列表 （第二个项目符号）。	第 233 页上 “ 使用 RDMA 的 NVMe-oF 配置 ”
澄清了 步骤 4：“ 要**通过 PowerShell** 启用 RDMA，请发出以下 Windows PowerShell 命令 ”。	第 250 页上 “ 创建带 RDMA NIC 的 Hyper-V 虚拟交换机 ”
在步骤 2 中，将 PowerShell 命令的最后一个字更正为 `-ManagementOS`。	第 251 页上 “ 将 vLAN ID 添加到主机虚拟 NIC”
更改了 c 部分 步骤 1 中的命令。添加了 `ovsdb-server` 和 `ovs-vswitchd` 与 `pid` 一起运行的示例。	第 244 页上 “ 在 Linux 上配置 VXLAN”
在 c 部分 步骤 4 中，将第二段第二句改为 “`br1` 接口命名为 eth0， ens7 ；**通过网络设备文件手动配置静态 Ip，并将相同的子网 IP 分配给对等机 （主机 2 虚拟机）。** ”	
更改了监视虚拟机上虚拟功能流量的 PowerShell 命令。	第 269 页上 “ 监测流量统计信息 ”
更改了有关配置群集见证的详细信息的链接。	第 274 页上 “ 步骤 3. 配置群集见证 ”
在第一段中，将 cmdlet 更改为 `Enable-ClusterS2D`。	第 275 页上 “ 步骤 5. 启用 Storage Spaces Direct”
在 表 A-1 中，在链路 LED 部分，更新了 LED 外观和网络状态列。	第 287 页上 “ 适配器 LED”

词汇表

传输控制协议
请参阅 TCP。

传输控制协议 / 互联网协议
请参阅 TCP/IP。

串行化器 / 并行化器
请参阅 SerDes。

大量发送卸载
请参阅 LSO。

带宽
以特定传输速率可传输的数据量的度量。1Gbps 或 2Gbps 光纤信道端口可传输或接收的标称速率为 1 或 2Gbps，具体视所连接的设备而定。这分别对应于实际带宽值 106MB 和 212MB。

单根输入 / 输出虚拟化
请参阅 SR-IOV。

第二层
指多层通信模型开放系统互连 (OSI) 的数据链路层。数据链路层的功能是跨网络中的物理链路移动数据，其中交换机使用目标 MAC 地址在第二层级别重定向数据消息以确定消息目标。

动态主机配置协议
请参阅 DHCP。

非易失性存储器表示
请参阅 NVMe。

非易失性随机存取存储器
请参阅 NVRAM。

服务质量
请参阅 QoS。

高级配置和电源接口
请参阅 ACPI。

互联网广域 RDMA 协议
请参阅 iWARP。

互联网小型计算机系统接口
请参阅 iSCSI。

互联网协议
请参阅 IP。

基本地址寄存器
请参阅 BAR。

基本输入输出系统
请参阅 BIOS。

基于聚合以太网的 RDMA
请参阅 RoCE。

节能以太网
请参阅 EEE。

精简指令集计算机
请参阅 RISC。

巨型帧

在高性能网络中使用以提高长距离性能的大型 IP 帧。对于千兆位以太网，巨型帧通常意味着 9,000 字节，但可指超出 IP MTU（在以太网上为 1,500 字节）的任意大小。

可扩展固件接口

请参阅 EFI。

类型长度值

请参阅 TLV。

链路层发现协议

请参阅 LLDP。

目标

SCSI 会话的存储设备端点。启动器从目标请求数据。目标通常为磁盘驱动器、磁带驱动器或其他媒体设备。通常， SCSI 外围设备是目标，但某些情况下，适配器也可能是目标。目标可包含许多 LUN。

目标是响应启动器 （主机系统）请求的设备。外围设备是目标，但对于某些命令（例如， SCSI COPY 命令）来说，外围设备可充当启动器。

驱动程序

用于充当文件系统与物理数据存储设备或网络介质之间的接口的软件。

人机界面基础设施

请参阅 HII。

设备

目标，通常为磁盘驱动器。安装或连接到系统的磁盘驱动器、磁带驱动器、打印机或键盘等硬件。在光纤信道中，为*目标设备*。

适配器

用于联接主机系统与目标设备的板卡。适配器的其他表示形式包括主机总线适配器、主机适配器和板。

适配器端口

适配器板上的端口。

数据中心桥接

请参阅 DCB。

数据中心桥接交换

请参阅 DCBX。

统一可扩展固件接口

请参阅 UEFI。

网络接口卡

请参阅 NIC。

文件传输协议

请参阅 FTP。

消息信号中断

请参阅 MSI， MSI-X。

小型计算机系统接口

请参阅 SCSI。

虚拟机

请参阅 VM。

虚拟接口

请参阅 VI。

虚拟逻辑区域网络

请参阅 vLAN。

以太网

最为广泛使用的 LAN 技术，用于在计算机之间发送信息，通常速度为每秒 10 和 100 兆比特 (Mbps)。

以太网光纤信道

请参阅 FCoE。

用户数据报协议

请参阅 UDP。

远程直接内存访问

请参阅 RDMA。

增强的传输选择

请参阅 ETS。

最大传输单元

请参阅 MTU。

质询 - 握手身份验证协议

请参阅 CHAP。

聚合网络适配器

Marvell 聚合网络适配器在使用两种新技术的单一 I/O 适配器上支持数据网络 (TCP/IP) 和存储网络 (光纤信道) 流量：增强型以太网和以太网光纤信道 (FCoE)。

光纤信道

支持其他较高层协议 （如 SCSI 和 IP）的高速串行接口技术。

主机

一个内存或复合 CPU 管理的一个或多个适配器。

主机总线适配器

用于将主机系统 （计算机）连接到其他网络和存储设备的适配器。

虚拟端口

请参阅 vPort。

ACPI

*高级配置和电源接口 (ACPI) 规格*提供了统一的、以操作系统为中心的设备配置和电源管理的开放标准。 ACPI 定义了与平台无关的接口，用于硬件查找、配置、电源管理和监测。该规格以操作系统导向的配置和电源管理 （OSPM，一个用于描述实施 ACPI 的系统的术语）为中心，从而移除了旧版固件接口的设备管理责任。

BAR

基本地址寄存器。用于保留设备所使用的内存地址，或端口地址的偏移。通常，内存地址 BAR 必须位于物理 RAM 中，而 I/O 空间 BAR 可以驻留在任意内存地址（即使超出物理内存范围）。

BIOS

基本输入输出系统。通常位于 Flash PROM 中，用于充当硬件和操作系统间接口的程序 （或公用程序），并且允许在启动时从适配器引导。

CHAP

质询 - 握手身份验证协议 (CHAP) 用于远程登录，通常发生在客户端与服务器或 Web 浏览器与 Web 服务器之间。质询 / 响应是一种安全机制，用于验证人员或过程的身份而不泄露两个实体共享的密码。也称为*三方握手*。

CNA

请参阅 聚合网络适配器。

DCB

数据中心桥接。提供对现有 802.1 桥接规范的增强，以满足数据中心内的协议和应用程序的需求。由于现有高性能数据中心通常包含在不同链路层技术上运行的多个应用特定网络 （光纤信道用于存储，以太网用于网络管理和 LAN 连接）， DCB 允许使用 802.1 桥接来部署聚合网络，从而使所有应用可以在单个物理基础结构上运行。

DCBX

数据中心桥接交换。 DCB 设备用来与直连对等方交换配置信息的协议。该协议也可用于误配置检测和对等方配置。

DHCP

动态主机配置协议。仅当请求后，允许 IP 网络上的计算机从具有该计算机相关信息的服务器提取其配置。

eCore

OS 与硬件和固件之间的层。它是设备特定的，与 OS 无关。 eCore 代码需要 OS 服务 （例如，进行内存分配、 PCI 配置空间访问等）时，它会调用在 OS 特定层中实施的抽象 OS 功能。 eCore 流可能通过硬件驱动 （例如，通过中断）或通过驱动程序的 OS 特定部分驱动 （例如，加载和卸载 load 和 unload）。

EEE

节能以太网。计算机网络标准的双绞线和背板以太网家族的一系列增强功能，可在低数据活动期间降低功耗。目的是将功耗降低 50% 或更多，同时保持与现有设备的完全兼容性。电气和电子工程师协会 (IEEE) 通过 IEEE 802.3az 特别工作组制定了该标准。

EFI

可扩展固件接口。此规范定义操作系统与平台固件之间的软件接口。EFI 取代了在所有 IBM PC 兼容个人计算机中提供的旧版 BIOS 固件接口。

ETS

增强的传输选择。用于指定传输选择增强的标准，以支持流量类型之间的带宽分配。当某个流量类型中提供的负载不使用为其分配的带宽时，增强的传输选择允许其他流量类型使用可用带宽。带宽分配优先级与严格优先级共存。 ETS 包含管理对象以支持带宽分配。有关更多信息，请参阅：http://ieee802.org/1/pages/802.1az.html

FCoE

以太网光纤信道。一种由 T11 标准机构定义的新技术，通过在第二层以太网帧内封装光纤信道帧，允许传统的光纤信道存储网络流量在以太网链路上传输。有关更多信息，请访问 www.fcoe.com。

FTP

文件传输协议。一种标准网络协议，用于将文件通过基于 TCP 的网络 （如互联网）从一台主机传送到另一台主机。进行带外固件上传 （比带内固件上传速度更快）时需使用 FTP。

HBA

请参阅 主机总线适配器。

HII

人机界面基础设施。此规范 （UEFI 2.1 的一部分）用于管理用户输入、本地化字符串、字体和窗体，允许 OEM 开发用于预引导配置的图形界面。

IEEE

电气和电子工程师协会。一个旨在促进电气相关技术进步的国际性非营利组织。

IP

互联网协议。一种通过互联网将数据从一台计算机发送到另一台计算机的方法。 IP 指定数据包的格式 （也称为*数据报*）和寻址方案。

IQN

iSCSI 限定名称。 iSCSI 节点名称基于启动器制造商和唯一的设备名称部分。

iSCSI

互联网小型计算机系统接口。一种将数据封装到 IP 数据包以通过以太网连接发送的协议。

iSCSI 限定名称

请参阅 IQN。

iWARP

互联网广域 RDMA 协议。一种实现 RDMA 的网络协议，用于通过 IP 网络进行有效的数据传输。 iWARP 设计用于多种环境，包括 LAN、存储网络、数据中心网络和 WAN。

LLDP

供应商中立的第 2 层协议，允许网络设备在本地网络上公布其身份和功能。该协议取代了专有协议，如思科发现协议、极端发现协议和 Nortel 发现协议 （也称为 SONMP）。

使用 LLDP 收集的信息存储在设备中，可以使用 SNMP 查询。通过抓取主机并查询该数据库，可以发现启用 LLDP 的网络的拓扑。

LSO

大量发送卸载。 LSO 以太网适配器功能，允许 TCP\IP 网络堆栈构建大型 （最多 64KB） TCP 消息，然后将其发送给适配器。适配器硬件将消息分段为可通过线路发送的较小数据包 （帧）：对于标准以太网帧最多为 1,500 字节，对于巨型以太网帧最多为 9,000 字节。分段过程可释放服务器 CPU，从而不必将大型 TCP 消息分段为装进所支持帧大小的较小数据包。

MSI， MSI-X

消息信号中断。两个由 PCI 定义的扩展之一，用于支持 PCI 2.2 及更高版本和 PCI Express 中的消息信号中断 (MSI)。 MSI 是另一种通过特殊消息来生成中断的方式，可模拟引脚有效或无效置位。

MSI-X （在 PCI 3.0 中定义）允许设备分配介于 1 到 2,048 之间任意数量的中断，并为每个中断提供单独的数据和地址寄存器。 MSI 中可选的功能 （64 位寻址和中断屏蔽）对 MSI-X 为强制的。

MTU

最大传输单元。是指指定通信层协议可以传输的最大数据包 （IP 数据报）大小 （以字节为单位）。

NIC

网络接口卡。安装以启用专用网络连接的计算机卡。

NIC 分区

请参阅 NPAR。

NPAR

NIC 分区。将一个 NIC 端口分成多个物理功能或分区，每个均有用户可配置的带宽和个性设置 （接口类型）。个性设置包括 NIC、 FCoE 和 iSCSI。

NVRAM

非易失性随机存取存储器。一种在即使断电时仍可保留数据 （配置设置）的存储器。您可以手动配置或从文件还原 NVRAM 设置。

NVMe

专为销售状态驱动程序 (SSD) 设计的存储访问方法。

OFED™

OpenFabrics 企业分布。 RDMA 和内核旁路应用程序的开源软件。

PCI™

外围组件接口。一种由 Intel® 引入的 32 位本地总线规格。

PCI Express (PCIe)

第三代 I/O 标准，除支持旧式外围组件互连 (PCI) 和 PCI 扩展 (PCI-X) 台式机和服务器插槽外，还支持增强型以太网网络性能。

QoS

服务质量。在通过虚拟端口传输数据时，此方法用于通过设置优先级和分配带宽来预防瓶颈和确保业务连续性。

PF

物理功能。

RDMA

远程直接内存访问。此功能允许一个节点通过网络直接写入另一个节点的内存 （使用地址和大小语义）。此功能是 VI 网络的一项重要功能。

RISC

精简指令集计算机。一种计算机微处理器，它执行较少类型的计算机指令，从而以更高的速度运行。

RoCE

基于聚合以太网的 RDMA。一种网络协议，允许通过聚合或非聚合以太网络进行远程直接内存访问 (RDMA)。RoCE 是链路层协议，允许位于同一以太网广播域中的任意两个主机之间进行通信。

SCSI

小型计算机系统接口。一种用于将硬盘驱动器、 CD 驱动器、打印机和扫描仪等设备连接到计算机的高速接口。 SCSI 可使用一个控制器连接许多设备。每个设备均可通过 SCSI 控制器总线上单独的标识号进行访问。

SerDes

串行化器 / 并行化器。一对功能块，通常在高速通信中用于弥补受限的输入 / 输出。这些功能块在串行数据接口和并行接口之间沿每个方向转换数据。

SR-IOV

单根输入 / 输出虚拟化。一种 PCI SIG 的规格，使单个 PCIe 设备能够显示为多个单独的物理 PCIe 设备。 SR-IOV 允许隔离 PCIe 资源以实现性能、互操作性和可管理性。

TCP

传输控制协议。一组通过互联网协议发送数据包中数据的规则。

TCP/IP

传输控制协议 / 互联网协议。互联网的基本通信语言。

TLV

类型长度值。可编码为协议内元素的可选信息。类型和长度字段为固定大小 （通常 1-4 字节），而值字段为可变大小。这些字段用法如下：

- 类型 - 表示此部分消息所代表字段类型的数字代码。
- 长度 - 值字段的大小 （通常以字节为单位）。
- 值 - 一组包含此部分消息数据的、可变大小的字节。

UDP

用户数据报协议。一种不保证数据包顺序或传递的无连接传输协议。它直接在 IP 之上工作。

UEFI

统一可扩展固件接口。此规范详细说明了一个接口，可帮助将预引导环境 （即从系统开机之后到操作系统启动之前）的系统控制移交给操作系统 （如 Windows 或 Linux）。在引导时， UEFI 提供操作系统与平台固件之间的干净接口，并支持用于初始化添加式卡的独立于体系结构的机制。

VF

虚拟功能。

VI

虚拟接口。用于跨光纤信道和其他通信协议进行远程直接内存访问的方案。用于群集和消息传递。

vLAN

虚拟逻辑区域网络 (LAN)。具有一组通用要求的主机组，如同连接到同一线路一样通信，无论其物理位置如何。尽管 vLAN 具有与物理 LAN 相同的属性，但它允许终端站组合在一起，即使其不在同一 LAN 网段上。vLAN 可通过软件重新配置网络，而不是物理搬动设备

VM

虚拟机。机器 （计算机）的软件实现，如同真实机器一样执行程序。

vPort

虚拟端口。与一个或多个虚拟服务器关联的端口号或服务名称。虚拟端口号应是客户端程序预期连接的同一个 TCP 或 UDP 端口号。

LAN 唤醒

请参阅 WoL。

WoL

LAN 唤醒。一种以太网计算机联网标准，允许通过发送的网络消息远程开启或唤醒计算机，这些消息通常由在网络中另一台计算机上执行的简单程序发送。

Marvell 科技集团
http://www.marvell.com

Marvell. 快步向前