倾情奉献

IBM BROCADE

Broadcom 子公司

NVMe over Fibre Channel

傻瓜书

Wiley

利用超低的延迟，提升性能

维持任务关键型存储SLA

利用并发SCSI/NVMe，降低风险

Brian Sherman

Marcus Thordal

Kip Hanson

IBM/Brocade

特别版第二版

关于 IBM

IBM 是一家全球性的技术与创新公司，总部位于纽约州阿蒙克市。IBM 不仅仅是一家“硬件、软件与服务”公司，如今它已发展成为一家认知解决方案和云平台公司。超过 25,000 家公司因 IBM 和 Brocade 在创新和专业知识方面的组合而选择了这两家公司的解决方案和服务。IBM 已帮助许多像您这样的企业改善了安全性、协作水平、生产效率和运营水平。这两家公司强强联手，可帮助您实现业务目标。

有关 IBM 闪存存储解决方案的更多信息，敬请访问：**ibm.biz/flashstorage。**

有关 IBM b-type SAN Storage 解决方案的更多信息，敬请访问：**ibm.biz/san-btype。**

关于 Brocade

Brocade（Broadcom Company 子公司之一）是光纤通道存储网络领域一家久经验证的领先公司，其光纤通道网络用于为虚拟化全闪存数据中心提供基础。自 1998 年起，Brocade 开始与 IBM 联手提供光纤通道存储解决方案，此类解决方案可交付创新型的高性能网络，这些网络不仅极具弹性，而且易于部署、管理和扩展，适于各种要求最苛刻的环境。网络对于存储而言非常重要，而 Brocade 的光纤通道存储网络解决方案可为企业存储提供最可信、部署最为广泛的网络基础架构。

www.broadcom .com

NVMe over Fibre Channel

IBM/Brocade 特别版第二版

作者：Brian Sherman、Marcus Thordal、Kip Hanson

傻瓜书

Wiley

NVMe Over Fibre Channel 傻瓜书，IBM/Brocade 特别版第二版

出版商：

John Wiley & Sons, Inc.

111 River St.

Hoboken, NJ 07030-5774

www.wiley.com

ISBN 978-1-119-60267-5 (pbk)；ISBN 978-1-119-60270-5 (ebk)

美国印刷

10 9 8 7 6 5 4 3 2 1

出版商鸣谢

对于本书及参与本书出版的人士，我们深感自豪。谨此感谢帮助本书成功出版的人士：

第一版联合作者：Curt Beckmann

项目编辑：Martin V. Minner

编辑经理：Rev Mengle

执行编辑：Katie Mohr

业务开发代表：Karen Hattan

制作编辑：Siddique Shaik

Brocade 一方的贡献人员：Marc Angelinovich、AJ Casamento、Howard Johnson、David Peterson、David Schmeichel

目录

引言

自*斯普特尼克 1 号*发射以来，只要您不是在西伯利亚隐居放牧驯鹿，您可能就知道现在小孩所用的旧 iPhone 3 不论是在计算能力还是容量上都已经秒杀了*阿波罗 13 号*。您可能已经意识到，创新的步伐仍在不断加快。如今不仅网速比以前快了数百万倍，每秒生成的数据也多了数百万倍。处理速度也呈指数级加快。就存储而言，整个美国国会图书馆的内容都可以保存在价格低廉的磁盘阵列中。时代真的变了。

关于本书

本书《NVMe Over Fibre Channel 傻瓜书》IBM/Brocade 特别版第二版聚焦的是信息技术中比较小但又很重要的方面。NVMe (Non-Volatile Memory Express) over Fibre Channel 是一种涉及计算机内存、存储和网络的技术。

如果您是一个资深电脑迷，对于这种技术您可能已经有所耳闻。如果不是，那么这次也不会是您最后一次听到这种技术。与 IT 领域的大多数技术一样，FC-NVMe (NVMe over Fibre Channel) 有一段丰富的历史，经历了高速的发展，它构建于以前的技术之上，同时规避了同类竞争技术的缺陷。假如您对该技术一无所知，那么在阅读本书后，您会发现 FC-NVMe 可能是网络存储领域的“明日之星”。

简而言之，NVMe over Fibre Channel 聚集了一切优势。它拥有工作内存应用所需的超低延迟，具有对企业存储来说至关重要的可靠性。网络极客都知道，光纤通道是一个主流的数据中心网络标准，因此，NVMe over Fibre Channel 能够利用基于结构的分区和名称服务。最重要的是，NVMe over Fibre Channel 完全遵守现有的光纤通道上层协议，让您能够以低风险的方式从 SCSI（Small Computer System Interface 的简称）迁移到 NVMe，而不需要投资实验性的基础架构。

假定事项

您的玛丽阿姨每周都会打电话要您帮她解决 Facebook 帐号问题，您对网络和存储技术的了解可能也比她多。如果您不了解网络和存储技术，本书提供了大量提醒和边栏，旨在帮助您掌握技术中的难点，并理解每个 IT 领域的各种令人困惑的缩写。

本书中使用的图标

《NVMe Over Fibre Channel 傻瓜书》中标记了一些有用的图标。这些图标将巩固并进一步解释重要的概念，帮助您满足上司的要求。

提示

特别注意“提示”图标。该图标代表该部分包含一小段有助于您简化工作的信息，避免因忙于在服务器机房内重新配置存储阵列而通宵加班。

切记

如果您整天在看厚厚的硬件手册，那么您肯定会在看书时忘记一些知识点。这时，“切记”图标可能是您的理想新伙伴。

警告

计算机硬件和网络设备价格昂贵。如果因为您决策失误而不得不更换设备，这又需要您投入更多成本。如果您想避免 IT 战略中那些代价高昂的错误，请留意“警告”图标。

技术内容

假如您是在格子间里悬挂技术大拿 Jack Kilby 和 Robert Metcalfe 照片的专业硬件专家，那么请注意“技术内容”图标；该图标代表本部分将针对一些深奥的主题额外提供详细的信息。

本章提要

- **传统存储与传统内存**
- **加速访问闪存**
- **NVMe 与 SCSI 之间有何种关系**
- **FC-NVMe 的未来优势**
- **激发结构的活力**

第 1 章

探索 NVMe over Fibre Channel

本章将提供一个有关 NVMe over Fibre Channel 的简明教程。我们将介绍（或者向专家重新介绍）一些令人困惑的缩写，探讨固态存储的优势，比较固态存储和内存，并对这项令人激动又相对较新的技术分析一下各个组成部分。我们希望您知道为什么 NVMe over Fibre Channel 可能是贵企业的最佳选择。

NVMe over Fibre Channel 是一项功能完备的高性能技术，面向的是基于 NVMe 且结构连接的企业存储，但它也是面向 NVMe 工作内存用例的理想解决方案。（我们将在本章中讨论这些用例有何差别。）NVMe over Fibre Channel 是一项比较新的技术，尽管其组成部分并非新技术。自 20 世纪 90 年代中期起，光纤通道 (FC) 就一直是领先的企业存储网络技术。速度达到 16 Gbps 的第五代光纤通道已经得到了广泛普及。2016 年第六代光纤通道问世，它将第五代光纤通道的速度提高了一倍，将 128GFC 链路上的带宽提高了 8 倍，目前第六代光纤通道正在热销。第七代光纤通道又会再次将速度提高一倍，随着第七代 HBA 的上市，第七代光纤通道也即将问世。FC 主要用于履行 SCSI 协议，它是面向直连式 PC 或服务器存储的传统领先技术。SCSI on Fibre Channel 也被大胆地称作光纤通道协议 (FCP)。

NVMe 与多个事项有关：

- 一个开放式标准集合，适用于访问和管理非易失性内存（NVM），尤其是诸如闪存一类的高性能固态内存或诸如 3D Xpoint 的存储级内存 (SCM)
- 该集合的主要规范，它提供通用的高性能接口，用于直接通过 PCI Express 访问 NVM（参见图 1-1）(https://nvmexpress.org/resources/specifications/)
- 非营利性组织 NVM Express (https://nvmexpress.org/)，该组织致力于在多家科技公司的支持下，制定和推广标准

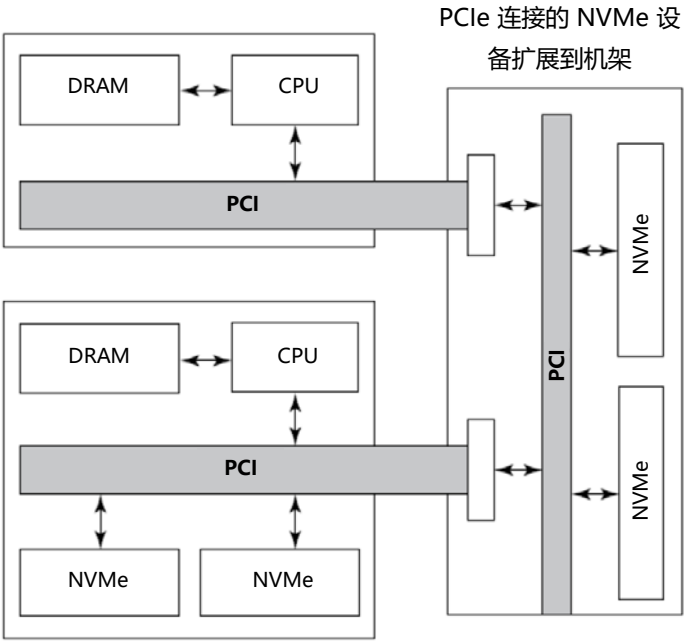


图 1-1：NVMe 在内部或外部连接服务器 PCIe 总线。

基于 PCI 的 NVMe 延迟低，但是相比基于结构的介质，它也有一些重要的限制。结构连接的优势包括共享访问、提高容量、增强数据保护和提供灵活的多供应商支持。使用结构还能帮助您消除单点故障，简化管理。为了让 NVMe 生态系统享受所有这些优势，NVM Express 创建了 NVMe over Fabrics（又名 NVMe-oF），从而定义如何以一致且与结构无关的方式在不同的结构之间传输 NVMe 命令（参见图 1-2）。这样，软件开发人员就能更轻松地开展工作了！

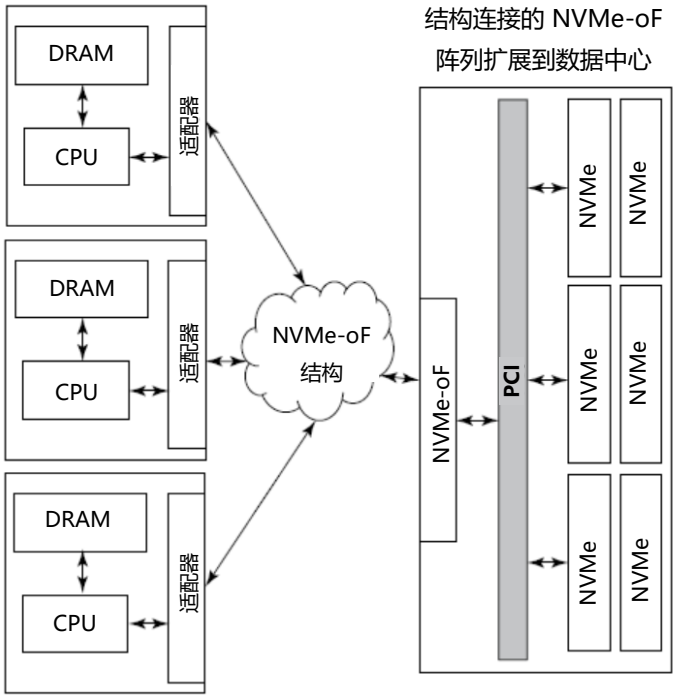


图 1-2：借助 NVMe over Fabrics，扩大 NVMe 的规模。

2016 年发布的 NVMe-oF 1.0 规范描述了两种结构类型，光纤通道结构和远程直接内存访问 (RDMA) 结构。

光纤通道结构：NVM Express 选择由处理所有光纤通道标准的 T11 标准组织来定义新的“FC-NVMe”协议。在将 NVMe 映射至光纤通道时，T11 委员会成员跟随 SCSI 的脚步，从而能够直接在同一基础架构上同时运行 SCSI 和 NVMe 流量。T11 委员会于 2017 年 10 月完成了这项工作。

RDMA 结构：RDMA 是一项与 InfiniBand、RoCE（读音与“rocky”相同）和 iWARP（我们提醒您注意缩写）有关的既定协议，现已实施多年。通过基于 RDMA 构建，NVM Express 可以在开展一项工作时同时瞄准三个现有的结构传输。

提示

2017 年初（NVMe-oF 1.0 发布后），NVM Express 的一个团队推进了一项工作，以映射 NVMe over TCP（没有 RDMA）。这样，即使没有 RDMA 的支持，NVMe-oF 也能在现有的数据中心内运行。

如需了解最新规范，敬请访问 https://nvmexpress.org/resources/specifications/。

切记

光纤通道能够并行传输多个更高级别的协议，比如 FCP 和 FC-NVMe（特定帧类型的 NVMe over Fibre Channel 流量标签）以及大型机存储协议 FICON。再重复一次：FC-NVMe 可以在您的 FC SAN 和 HBA 上与您现有的 FCP 或 FICON 流量并存。

NVMe over Fibre Channel 提供强大的可互操作性、超快的性能和高度可扩展的架构。不论您是需要升级传统的存储网络，还是实施以内存为中心的全新产品，NVMe over Fibre Channel 都能提供两全其美的解决方案，同时支持传统用户平稳完成转变。

站在选择方的立场：它是存储、网络还是内存?

有些人可能发现 *NVMe over Fibre Channel* 一词中有些自相矛盾之处。那是因为 FC 是以存储为导向的技术，而 *NVM* 一词又显然是以内存为导向。其他三个 NVMe 结构（InfiniBand、RoCE 和 iWARP）以内存为导向（它们支持远程直接内存访问，也即 RDMA），而 NVMe/ TCP 属于传统的网络访问。事实上，近期召开的有关闪存和其他持久内存技术的会议已经开始热烈地讨论起“存储/内存融合”的趋势。

映射两级

等一下……内存和存储正在融合？*什么？* 几十年来，内存和存储一直站在对立的两端。两者都能保存信息，但是内存植入在服务器内，而存储大多都是单独存在的，不依赖服务器或应用来保存数据。在某种程度上，这种两极化得到了自我强化：

» 与动态随机访问内存 (DRAM) 相比，传统的企业存储相对更慢，其错误检查和读/写往往是按顺序进行的。内存则速度很快但只是临时保存信息。

» 硬盘驱动器 (HDD) 和固态硬盘 (SSD) 绝不仅限于常规的内存。它们的单位比特成本更低，在断电时也能持久保存数据，这一点对于归档至关重要。

» 企业存储还支持各种服务级别协议 (SLA)，并提供强大的功能，比如独立磁盘冗余阵列 (RAID)、复制、去重和压缩功能。试试内存的这些功能。

表 1-1 比较了内存和存储的特点。

表 1-1 **内存和存储的特点**

功能	“理想内存”的优先级	闪存就像.....	NVMe 协议瞄准的是.....	“理想存储”的优先级
读取带宽	很高	内存	内存	中等
写入带宽	很高	存储	内存	中等
读取延迟	很高	内存	50/50	中等
写入延迟	很高	存储	50/50	中等
读取粒度	高	内存	存储	低
写入粒度	高	存储	存储	低
规模	GB 到 TB	GB 到 PB	存储	TB 到 EB
随机访问	很高	内存	内存	低
持久性	低	存储	存储	很高
可重写性	高	存储	两者	低到中等
可靠性	高	内存	存储	很高
密度	中等	存储	存储	高

因此，内存和存储依然看起来不同，并且这种现状可能会持续。借用马克吐温的话说，有关内存和存储融合的报告可能有些夸张。我们可以这样说，融合成为了一种趋势，而非迫在眉睫的事件。我们也可以说，共享内存协议和共享存储协议正在融合。

对于错误的态度

在计算中内存的错误大多被包容（或者至少没有被删除），而存储的错误则不然，这一点尽管可以理解，但是多少有点讽刺意味。这种情况在很多层面都存在。笔记本电脑（用户级别的计算）通常没有针对 DRAM 的纠错功能，但是驱动器中内置了 CRC 错误检测功能。服务器（企业级别的计算）有 ECC DRAM，后者会纠正单比特错误，但是出现双比特错误时则只能中止或关闭。相反，企业级存储包含某种形式的冗余，比如 RAID 或纠删码。

行业所采用的方法是利用保存的相同的数据中止并重启计算，而非修复每个内存错误。这是因为存储提供更高级别的保证或服务级别协议，即，存储客户与存储提供商之间的合同。存储消费者的组织及其 IT 部门经常使用这一术语。

警告

解决工作内存的问题并不能完全解决罕见但又意义重大的计算不可信问题。很多事件都会中断正在进行的计算，病毒、网络连接丢失和停电还只是冰山一角。正因为此，过度投资工作内存没有多大的意义。相反，由于没有“重做”机制，长期存储必须是可恢复的。结论呢？您在存储上千万别省钱。我们将在后面的章节中详细讨论这一点。

有关数据的其他数据

数十年来，主流的存储技术一直是以磁盘和磁带形式存在的磁记录技术，而主流的内存技术则是硅技术（大多为 DRAM）。闪存密度和性能的稳步提升是 NVMe 协议背后的主要技术推动因素。闪存早在多年前就已取代基于磁盘的存储；过去十年闪存成为了笔记本电脑的默认存储方式，自 2015 年起，闪存又发展成 NVMe 连接的 SSD。

加速访问闪存

闪存彻底颠覆了存储市场，但是在存储行业，这并非第一次。早在 1995 年，存储市场就成为了颠覆市场格局的榜样，哈佛大学的 Clayton Christensen 在其经典的哈佛商学院颠覆报告中引用存储市场作为主要案例。早期的颠覆更多的是与规模和成本有关，而闪存则更多的是与性能有关。

相比旋转磁盘驱动器，闪存始终能提供更快速的读取功能（尤其是针对随机访问提供读取功能）。但是早期闪存的密度比磁盘驱动器更低。此外，写入闪存比写入 DRAM 或磁存储介质难度更大。闪存拥有相对更低的写耐久性，每个闪存块的擦除/写周期大约为 100 万次。对一个闪存块的重复写入也会降低相邻闪存块的可靠性。因此，尽管闪存的速度很快，但是其早期缺陷限制了它只能用于某些小众用途。

随着时间的推移，闪存的密度大幅提高，同时，企业也开发出了有效的软件算法来解决写入挑战。凭借速度、密度和可接受的写耐久性，闪存成为了企业保存生产数据的理想技术，取代了数据中心里的旋转磁盘。

事实上，闪存以及其他固态内存技术的超高速度优势证明了久经考验的传统存储协议有一个缺陷：性能。

了解 NVMe 与 SCSI 之间有何种关系

当今大多数以存储为导向的协议（包括 FCP）的基础都是 20 世纪 80 年代建立的 SCSI 标准。SCSI 一开始是围绕硬盘驱动器构建的，后来经过多次扩展才涵盖其他存储设备，同时保持向后兼容性。SCSI 目前支持 100 多种命令。除了承担了很多负担外，SCSI 还缺乏深度命令队列。

SCSI 为传统应用提供的无数扩展组件和扩展支持衍生出了一个协议栈，它比 NVMe 协议栈更迟钝，而 NVMe 协议栈拥有大幅增强的队列，并针对半导体内存和当今的操作系统进行了简化和优化（参见图 1-3）。

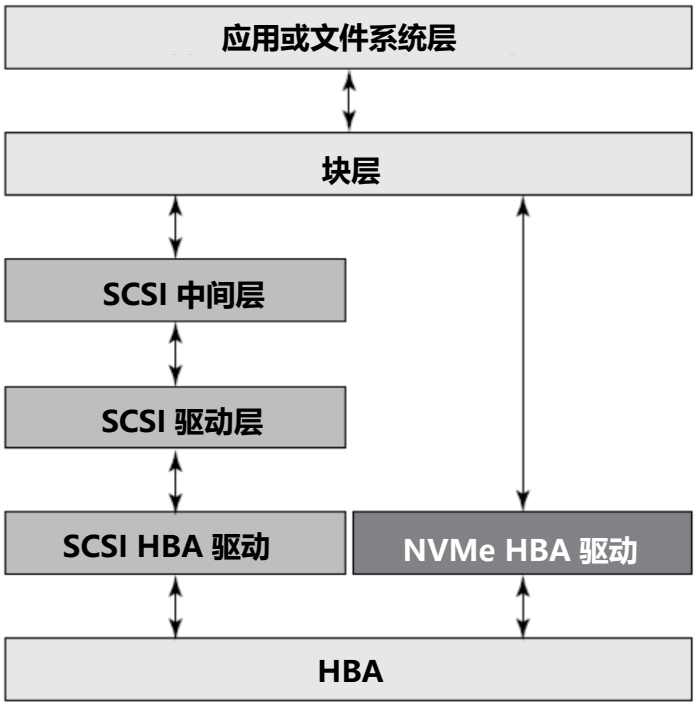


图 1-3： 比较 SCSI 和 NVMe 软件堆栈。

假如您可能在某个时候将基于 SCSI 的存储资产迁移到 NVMe 环境，那么您需要知道以下重要信息：

- **将传统 SCSI 映射到 NVMe：** NVMe 社区已经意识到了存储市场的重要性以及 SCSI 在存储市场的突出地位。正因为此，NVMe 标准组织投入了大量时间和精力，确保 NVMe 能够实施依赖传统存储的应用所需的功能。
- **LUN 和命名空间 ID：** LUN 代表的是逻辑单元号，这是一个 SCSI 机制，用于识别单一存储资产内不同的卷。换句话说，每一个卷就是一个 LUN。NVMe 以类似的方式使用术语 *命名空间 ID (NSID)*。*命名空间*是一个奇妙的术语，因为每个命名空间都被当作一组逻辑块地址 (LBA)，而非一组名称。

- **增强型命令队列**：FC-NVMe 提供 NVMe 的增强型队列功能，允许通过单一连接传输数千个并行请求。如今的多线程服务器和虚拟机运行了数十个甚至数百个应用，这种情况下，并行性将带来巨大的优势。
- **光纤通道协议（FCP）**：与*命名空间*一样，*光纤通道协议*也是一个奇怪的名字，它与“SCSI over Fibre Channel”非常类似。FCP 与基础光纤通道无关；它与在光纤通道上实施 SCSI 功能的方式有关。
- **利用 FCP**：即使没有详细的信息，您也应该知道 NVMe over Fibre Channel 采用了全新的 FC-NVMe 帧类型处理非 I/O 通信，同时重复使用 FCP 帧类型处理 I/O 操作。因此，如果您捕获了所有跨 NVMe-over-FC 接口运行的帧，您就会看到其中有 FCP。
- **其他协议**：人们之所以认为光纤通道仅适用于 SCSI，*FCP for SCSI-on-FC* 这个名字可能要承担一部分责任，但是切记一点，SCSI 并非唯一一个使用 FC 的热门协议。大型机存储协议 FICON 也在 FC 上运行，NVMe 则是另一个在 FC 上运行的协议（该协议是本书的讨论重点）。

警告

SCSI 的迟缓是协议栈的特点，而非光纤通道传输的特点。一些 NVMe 支持者转而实施 SCSI over FC，他们将性能相对缓慢的责任错误地归咎于光纤通道。这是一种错误的主张。FC-NVMe 的速度比 SCSI over FC 快。

预测 NVMe over Fibre Channel 的未来优势

可扩展性是 NVMe over Fibre Channel 的重要优势之一。它围绕非易失性内存从头构建而成，同时也结合了光纤通道的速度和稳健性。（我们将在第 2 章更详细地介绍光纤通道的固有优势。）通过利用光纤通道进行传输，用户能够轻松享受 NVMe over Fabrics 的所有速度和并行优势，同时又能够规避构建并行基础架构所带来的业务中断。

随着 NVMe over Fibre Channel 的竞争优势越来越大，您还需要考虑以下因素：

- 随着闪存变得更加以存储为导向，主流存储协议（SCSI）已经削弱了闪存的一大优势：速度。随着越来越多的存储供应商开始拥抱 NVMe over Fibre Channel，这种局面将发生变化。
- Intel/Micron 的 3D Xpoint 这类新的半导体内存正在登上舞台。它们让我们看到了提高写入速度、以及将写耐久性提高 100 倍甚至 1000 倍的希望。

激发结构的活力

存储阵列的基础技术从旋转磁盘发展至闪存，又从闪存发展至速度更快的技术。这种情况下，速度的加快更加大了对于 NVMe 比 SCSI 多提供的宝贵数百微秒的需求。

此外，很多应用会有各种各样的需求，这又需要一些以存储为导向的卷和一些以内存为导向的卷。更有意思的是，有时您会想用相同的信息完成两项不同的工作。换言之，您希望维持一些数据资产的主副本，从而启用企业存储的所有高度可靠的功能。

与此同时，其他数据资产消费者可能只需要高速读取访问这些数据。您可以考虑在 NVMe over Fibre Channel 驱动上，（使用光纤通道结构的双协议并发功能）将您的主存储卷发布到“工作参考内存”。这类映像可能是只读的，它们不需要会产生延迟或成本的功能。通过使用以内存为导向的 NVMe over Fibre Channel 阵列，您也可以节省成本。

本章提要

- **审视光纤通道在存储生态系统中的地位**
- **在存储环境中评估性能指标**
- **认识改进型队列的优势**
- **实现高性能**
- **实现可靠性**

第 2 章

通过 NVMe over Fibre Channel 交付速度和可靠性

1997 年光纤通道问世时，以太网已经成为了主流网络选择。以太网的发展已经超越令牌环和异步传输模式 (ATM) 这类协议，很多网络社区的成员纷纷质疑 FC 是否还有未来。

那么，他们错了吗？面对以太网的强烈冲击，光纤通道网络成功迎难而上，发展到现在，几乎所有企业都依赖光纤通道来满足其任务关键型存储需求。光纤通道的成功并不神奇。FC 在很多重要的方面与以太网存在差异，并且现在依然如此。了解为什么光纤通道能够在以太网主宰的世界取得如此大的成功，这是您在决定沿用或者采用这项强大的网络技术时不可回避的第一步。

审视光纤通道在存储生态系统中的地位

以太网依然是负责服务器之间的通信的主要传输技术。但是，在以存储为中心的环境中使用光纤通道能为您带来明显的优势。传统以太网，包括现在部署的大多数以太网，并未采取多少措施来解决网络拥塞问题。相反，它将可靠传输的责任推给了上层协议。我们假设您的网络是处于高峰期的高速公路，以太网会允许无数车辆（帧）进入匝道，但是当高速公路上拥堵不堪时，以太网会将车辆引向沟里（丢帧）。即使车辆没有到达终点也不要紧：传输控制协议（TCP）会耐心又迟疑不定地再次尝试，根据需要发出所需数量的车辆，即使前方已经出现了车祸。

另一方面，光纤通道旨在以可靠且井然有序地方式传输数据，不论负载情况如何。它有一个功能就像我们每天通勤时都会抱怨的入口匝道信号灯。不会丢帧，它会妥善地交付每一帧。光纤通道高速公路上没有堵车也没有车祸。这也让它成为了企业满足任务关键型存储要求的理想解决方案。

在存储环境中评估性能指标

对于自身尚且无法度量的事物，您不可能将其发展壮大。正因为此，您在启动任何改进项目之前都必须先确定性能指标，这一点至关重要。即使是园艺这样的小事，假如种子埋得不够深或者用错了肥料，您可能最终也会颗粒无收。

存储也不例外。如果不能清楚地知道贵企业花一百万美元投资的硬件运转情况如何，或者数据是否丢失，亦或者用户是否在抱怨访问企业数据的速度过于缓慢，最终您可能还需要投入更多的时间。或许您可以从事园艺工作。

本部分聚焦存储指标：最终存储消费者最关心这些指标，而且存储指标还能帮助您更好地横向对比不同的 NVMe over Fabrics 选项。（如需简要了解结构指标，请参见下方边栏“结构指标”。）存储社区早已使用以下三大指标来衡量存储性能：

- 延迟
- 吞吐量
- 每秒输入/输出操作 (IOPS)

结构指标

结构指标往往与存储指标类似，如果您对其具体含义不了解，有时您可能会将两者混淆。不论是存储从业者还是网络从业者都应该特别注意两者之间的微妙差别，避免出现用苹果与橘子进行比较这样尴尬的局面。

存储延迟会衡量从始至终的所有存储操作，而结构延迟则会告诉您相对于直接连接，结构设备会给连接增加多少增量延迟。衡量的时间段为从帧的第一个字节进入到该字节首次传输的时间（“先进先出”——FIFO 模式）。结构延迟对用户读取和写入操作以及不同的 I/O 规模是一样的。

通常，结构吞吐量指的是当所有端口都以最大速度运行时通过结构的数据量。A64 端口，10G 设备通过吞吐量为 640 Gbps（如果输入和输出单独计算，吞吐量还将翻倍）。但是一些低端设备内部可能“过度订阅”，无法同时在所有端口上全速运行，因此，请注意查看细则。

结构指标与 IOPS 不存在任何对应关系。但是，我们也有 IOPS 指标，因为当 I/O 操作重叠时延迟指标无法全面反映真实情况。同样的，当网络中多股流量重叠引发拥塞时，没有哪个简单的指标可以成功捕获行为。

这些性能指标的相对重要性很大程度上取决于用户应用。相比其他指标，系统更注重将响应时间值延迟降至最低。流媒体播放高清视频会生成大量数据，因此，极高的吞吐量至为关键，而数据库中密集的读/写活动也需要大量 IOPS（读作 *eye-ops*）。甚至专家之间的观点也不统一。有些专家认为延迟是最重要的存储指标，而其他专家则认为 IOPS 是重中之重。

当然，除了网速外，总体性能中还有很多指标。如果您的硬盘驱动器是 dog，那为什么要用光纤通道呢？同样的，如果服务器依然在使用 Pentium Pro 处理器，那么您在进行任何网络升级之前，最好先解决服务器性能。正如生命中的一切，有些事情比其他事情更重要。毫无疑问，整个 NVMe 对话本身是从高速固态硬盘入手。

存储指标

下面，我们将简要介绍重要的存储指标：

- **延迟，**尤其是读取延迟是基于闪存的系统的主要优势。在执行基于 NVMe 的数据传输时，这是被宣扬次数最多的优势（通常与 SAS 或序列连接的 SCSI 对比）。

  存储用户关心总体操作延迟，即，从读取或写入操作开始到操作全部完成的时间。存储延迟取决于 I/O 操作的规模和方向，I/O 是随机进行还是按顺序进行，以及连接速度。在提到特定的延迟指标值时，您需要加入相关的值。

- **吞吐量**描述了存储设备读取或写入数据的速度（MB/秒或者 GB/秒）。这些指标更适用于大型 I/O 操作。与延迟衡量结果一样，吞吐量衡量结果应该指明 I/O 方向（读还是写）以及访问类型（顺序访问还是随机访问）。为了确保衡量结果的完整性，您可以加入 I/O 规模，但是如果没有提到 I/O 规模，您可以假设该指标适用于更大型的 I/O 操作。

- **IOPS** 将告诉您设备每秒能够处理多少单独的读取和/或写入操作。

  与延迟和吞吐量类似，IOPS 指标也会因为 I/O 的规模、方向和访问类型（顺序访问还是随机访问）的不同而不同。

通常，一台设备的 IOPS 指标在处理小型 I/O 操作时更高。正因为此，引用的 IOPS 指标通常适用于小型 I/O，比如 4 KB。但是，很多需要高 IOPS 的应用使用了更大型的 I/O，比如 64 KB。您需要谨慎地确保您的设备 IOPS 指标与应用的需求保持一致。

举一个简单的例子，IOPS 与延迟密切相关。假设您的连接能够在 1 毫秒内完成 4 KB 的读取操作，那么您可能会期望在 1 秒钟内执行 1000 次这样的 I/O 操作。事实上，情况可能就会是这样。但是实际情况并不一定都这么简单（本章后面将详述）。因此，您最好是同时参考延迟和 IOPS 指标。

警告

存储指标并不能帮助您掌握所有情况。为了进行横向对比，我们在可控的情境下进行了性能基准评估，比如将单个测试器连接单个设备。结果表明，在真实世界中，“每个人的状况可能不同。”

尽管在“可控情境”下进行性能基准评估是合理的，但是自然而然地就会衍生出压力，促使您调优设备，让它们在测试情境下表现出色，虽然在相似的真实环境下设备可能无法一直表现这么出色：

- **示例 1**：有些设备在顺序访问时，拥有出色的吞吐量或 IOPS 性能，但是这种性能可能只能在以下情境中实现：您只有少量存储客户发起操作。请求者数量太多会导致设备资源负载过重，进而丢失顺序性能优势。
- **示例 2**：有些闪存阵列会保存擦除的闪存块池，从而提高写入速度。在执行写入操作时，控制器将相关的逻辑块地址“重新映射”到池中的闪存块，在其中写入新数据，然后在后台标记待擦除的老闪存块。如果设备在擦除的闪存块上的运行效率低，“垃圾回收”进程可能会进入前台，进而大幅减慢正常操作的速度，直至进程结束。

因为真实世界并非可控的情境，想要一致高性能的 IT 架构师应确保他们的环境中拥有相应的工具来快速、深入地调查系统行为。IBM 和 Brocade（Broadcom Inc. Company 子公司之一）早就意识到，（与偏重商品的以太网市场不同）光纤通道客户的期望证明了投资分析工具的价值。假如客户想要享受 NVMe 技术的性能优势，他们就很有可能发现自己需要优化环境的工具。（如需了解更多信息，请参见边栏“Fabric Vision 功能”。）

Fabric Vision 功能

IBM b-type SAN 产品组合与 Fabric Vision 技术提供了多种工具，用于分析和提高 FC 结构性能与可靠性。下面，我们将概述部分重要功能。

IO Insight：IO Insight 能够在支持的产品上，通过集成式网络传感器以非侵入的方式主动监控存储设备的 IO 延迟和行为，进而提供有关问题的深入洞察力，并确保服务级别。

VM Insight：利用基于标准的端到端虚拟机 (VM) 标记功能，在整个存储结构中无缝监控虚拟机性能。管理员能够迅速确定 VM/应用性能异常的根源，基于 VM/应用要求配置并调整基础架构，实现服务级别目标。

Monitoring and Alerting Policy Suite (MAPS)：利用 MAPS 内基于规则/策略的预置模板，简化整个结构的阈值配置、监控和警报。管理员可以利用通用规则和策略一次性配置整个结构（或多个结构），或者针对特定端口或交换机元素自定义策略。借助 Flow Vision 和 VM Insight，管理员能够在 MAPS 策略中设置 VM 流程指标的阈值，从而在 VM 性能降级时接收警报通知。

Flow Vision：这是一组以流程为导向的工具，它们能帮助管理员识别、监控和分析特定应用与数据流，从而简化故障排除，将性能最大化，规避拥塞，并优化资源。下面列出了两个 Flow Vision 工具。

Flow Monitor：针对结构内流程提供全面的可视性，包括支持您自动学习流程，并以不中断业务的方式监控流程性能。

Flow Mirroring：支持您以不中断业务的方式创建特定应用和数据流或帧类型的副本，让您能够捕获这些副本用于深入分析。

提升设备指标

当随机访问数据和顺序访问数据同时发生时，*读缓存*能为硬盘驱动器提供一臂之力。这是因为您可以快速访问缓存中的数据用于随机读取，而硬盘驱动器能够读取整个磁盘道，缓存其他数据块用于顺序读取。尽管读缓存无法大幅提高 SSD 性能，但是它依然很重要，因为缓存能提供比 SSD 更短的响应时间。

*写缓存*能从两个角度为闪存提供帮助。

- **写入速度**：写入 DRAM 比写入闪存设备更快，因为在写入闪存设备之前必须先擦除闪存。
- **写耐久性**：应用可能在很短的时间内多次写入同一闪存块。写缓存可以稍等片刻，然后将多个缓存写入转变成单个闪存写入。

*并行性*指的是利用大量基础设备交付更高的吞吐量，以及有时也会交付更高的 IOPS。

*流水线*描述了一个并行运行多个功能的系统。比如，读取操作可能会分解为几个不同的功能阶段：命令预处理，物理访问后端设备，错误纠正以及发送。这些功能阶段就像一条流水线的不同部分，您会发现读取操作“穿过整个流水线”。通过流水线作业，系统能够同时实施两个读取操作的不同阶段。在进行读取和写入操作时，通常会使用不同的流水线。

通过流水线作业，设备的 IOPS 指标将超过您对其延迟指标的预期。比如，一个设备在执行 256 KB 的读取时延迟 1 毫秒，您可能预期该设备的 IOPS 指标为 1000。但是，一旦读取重叠（前一批读取还未完成后一批读取就已经发出），设备在执行 256 KB 的读取时可能交付 1500 的 IOPS 指标。

“单线程”应用可能很难执行重叠的读取或写入，但是大多数性能敏感型应用现在都是“多线程”应用，能够生成重叠的 I/O，并充分利用流水线型设备性能。此外，并行运行多个应用的虚拟化服务器也会发起多个重叠的 I/O。（这会产生一个有意思的副作用：微小的延迟变更不会改变整体 IOPS，具体取决于系统的瓶颈在哪里。我们将在本章后面详细讨论这一点。）

正如您所看到的，NVMe 标准包括能大幅加速重叠的 I/O 的架构增强包。

提示

如果您正考虑使用附带写缓存功能的设备，请确保电源故障写回行为与应用需求保持一致。如果应用要求所有写入操作都保持一致，那么设备必须保证停电时保存缓存内容。

警告

请注意，架构性能增强包只能取得这样的成效。写缓存能够为处理写突发提供帮助，但是如果长期请求的写入吞吐量超过了缓存背后硬件的承受能力，缓存会被填满，请求的写入操作将被节流，以匹配基础设备的吞吐量。当同一基础设备上的操作重叠时，流水线可能会丧失性能优势。您需要利用您的应用测试特定的产品。

实现高性能

与光纤通道一样，FC-NVMe 能够相对直接地提供高性能。光纤通道供应商从一开始就在全力提高性能优势，利用一流的速度和功能（如直接安置存储有效负载），减少内存复制开销。客户也做出了贡献，因为他们选择光纤连接的频率比主流网络更高。光纤通道利用单一网络层、简化的寻址和与拓扑结构无关的路由，后者支持并行使用所有链路，最终提供更简单的网络堆栈。此外，光纤通道结构通常被部署成并行、冗余且主动-主动的结构，这样，它不仅能提供可靠性，还能提供更高的性能。同样的，光纤通道内置的基于信用的流控功能不仅能提供可靠性，还能改进性能。

所有这些优化成果对于数据中心架构都有着非同寻常的意义，即使某些优化成果可能不太适用于校园或互联网环境。FC-NVMe 充分发挥了光纤通道的所有传统优势，同时提供了 NVMe 固有的其他性能优势。简化的协议栈就是我们前面提到过的一个优势。增强型队列是另一个重要的改进成果。

采用增强型队列

几十年来，存储供应商一直在展开激烈的竞争，以期在本章描述的三大指标上取得领导地位，他们早就开始使用缓存、并行性和流水线这类技术，改进性能指标。

我们前面探讨了为什么闪存驱动在延迟方面优于磁盘驱动，因为闪存驱动不需要坐等存储介质旋转或者读/写头缓慢地移动到相应的磁道。此外，闪存芯片比最小的磁盘驱动都要小得多，这是闪存驱动的另一个优势。这意味着，并行使用数千个闪存芯片比使用数千个磁盘驱动要简单得多。

与此同时，存储启动程序与存储目标一样，也开始向极端并行性过渡。服务器运行的线程、内核和虚拟机数量越来越多。结果就是，由此生成的并行 I/O 数量也陡增。

基于 SCSI 的设备提供并行性功能，允许存储启动程序并行将多条命令“放入队列”。但是单个 LUN（逻辑单元号或卷）和目标端口（通常支持多个 LUN）的 SCSI 队列深度都有所限制。每个 LUN 的 SCSI 队列深度限制在 8 到 32 条命令不等，每个端口最高不超过 512 条命令。过去，这种方式似乎足以胜任，但是随着当今环境的不断向外扩展，SCSI 已经开始感到步履维艰。

NVMe 的设计人员已经意识到了这些趋势，并相应地定义了该协议的队列深度。NVMe 支持 64 个队列，每个队列深度为 64000 条命令。


切记

增强型队列能够大幅提高并行性。这并不代表，您会马上看到 NVMe 性能提升 100 倍，但是我们不难想象 NVMe 性能将提升 2 倍甚至更多。您还可以预见 CPU 利用率大幅降低，因为 CPU 不用再坐等队列中的 IO 操作完成。借助 IBM b-type Fibre Channel 结构提供的高级分析功能，您可以在您的环境内跨越所有设备和应用跟踪队列深度。

实现可靠性

每个人都想要可靠性：可靠的汽车、可靠的员工和可靠的互联网。但是如果没有可靠的数据，您很难获得上述可靠的示例。

当然，所有 IT 从业者都知道，用户想要可靠的计算。可是，处理错误的方法多种多样，有些方法比其他方法花费的成本更高。相比投入资金（和电池寿命）在纠错电路（ECC）以最小化比特错误上，企业对笔记本电脑的偶尔崩溃忍受度更高。但是笔记本电脑有很多风险敞口，因此，用户会想方设法保护电脑，比如后台备份软件。


切记

很多企业的服务器都使用 ECC 来修复单比特错误，但是一旦出现双比特错误，服务器就崩溃了。服务器（比如笔记本电脑）也很容易受到病毒和停电的冲击，因此，企业不需要在修复双比特错误的杀手级 ECC 上投入更多。他们可以忍受意外的服务器崩溃，因为他们能够重新运行应用，获取结果。这仅仅是因为他们的企业存储能够保证关键数据资产的重要副本始终可用。如果您不能依靠这些资产，情况会有什么不同?

冗余网络和多路径 IO

*阿波罗 13 号*升空时，它安装了很多冗余系统，以确保在出现孤立的故障时，飞船依然能够正常运行。您的数据资产可能不像宇航员的生命那样珍贵，但是，您也有重要的资产，您也有足够的经验知道冗余能够在哪些领域发挥作用。对于企业存储客户来说，这事关他们的经济效益；适当数量的冗余有助于您实现 99.999% 甚至更高的可靠性，确保客户满意而归。

投机取巧只会让您因小失大，因为一旦客户失望，他们就会转投竞争对手的怀抱。


提示

企业存储供应商清楚这一点。他们重金投资研发，在存储目标和存储网络领域构建强大的系统，确保一旦罕见的故障不可避免，系统仍然能够正常运转。（该技术比 *Saturn V* 更成熟；遗憾的是，故障也更罕见。）

供应商还竭尽全力提供多路径 I/O 以确保您不会碰到单点故障；此外，服务升级窗口很小甚至没有，这也能为您带来额外的优势。客户用他们的钱包投票，有效地驱使存储供应商做到了这一点。供应商必须做好准备，针对他们自己的产品以及他们销售的其他产品提供全天候 (24/7) 的企业支持。

企业存储供应商了解他们所在的市场，即使是在 20 世纪 90 年代中期以太网市场份额遥遥领先其他协议的时候也不例外。尽管以太网很有吸引力，但供应商会出于很多原因选择光纤通道。

无耗网络的功能

丢了狗，然后车钥匙也丢了。珍视的东西丢失，这是很可怕的事情。因此，您必须保管好您珍视的东西。而纸吸管这类东西则不需要那么重视。这类东西就算丢了，您可以用新的来替代。

光纤通道一直是一项无耗网络技术。它珍视有效负载。每个光纤通道都由缓冲区信用阀值负责管控，接收者将与发送者共享这些缓冲区信用阀值。发送者知道接收端还有多少可用的缓冲区，他们不会发送接收者无法处理的帧。相反，几十年来，以太网一直属于有耗网络，它们像对待吸管一样对待数据包，它们会在各种各样的情况下丢包，并且它们依赖 TCP 或其他机制来替代丢失的“吸管”。数据中心桥接 (DCB) 是一个以太网变体，它使用 PAUSE 帧（而非缓冲区信用阀值）来避免丢包，但是 DCB 依然面临一些重要的可互操作性挑战。

安全性

显然，如果您珍视您的帧，您就需要保护它们，确保它们不被不当的人工操作（不论是简单的错误还是更麻烦的问题）所影响。光纤通道以多种方式提供额外的安全性。在确保全球最重要的数据的安全性方面，光纤通道已经获得了信任，光纤通道将这种安全模式融入了 NVMe over Fabrics 中。

切记

光纤通道的其中一项重要优势源自其专门化的特性。以数据中心为中心的光纤通道并非互联网协议。因此，黑客无法在互联网中推着光纤通道的帧穿过您的防火墙，进入您的数据中心。

光纤通道还提供分区服务，后者将存储访问控制集成到网络中。这种久经验证的服务对于所有企业存储供应商来说都奏效，甚至在多供应商环境中也非常有用。

好工具至关重要

人非圣贤，孰能无过。聪明的 IT 管理员知道用强大的工具来减少人为失误，这有助于他们职业生涯的发展。如需了解更多信息，请参见边栏“SAN 自动化与存储集成”。

SAN 自动化与存储集成

IBM b-type SAN 自动化产品利用智能自动化功能，实现简单、无误的 SAN 配置和管理。很多企业阵列都能自动集成存储配置和 SAN 分区服务，通过单一管理点交付端到端的配置。这些功能也适用于 FC-NVMe。

本章提要

- **确定您的现状**
- **考虑您的采用战略**
- **充分利用双协议 FCP 和 FC-NVMe**
- **熟悉 NVMe over Fibre Channel**

第 3 章

采用和部署 FC-NVMe

您已经做完了所有“家庭作业”。您已经阅读了很多手册，参加了数次研讨会，与同事也进行了讨论。团队所有成员都认为是时候在企业内进一步采用 NVMe over Fibre Channel 了。那么，还有一个大问题：从何处入手?

确定您的现状

在您实施 NVMe over Fibre Channel 时，有几个因素对您有利。您不需要拿出部分预算，投资并行基础架构。您不需要担心设备的多供应商可互操作性，也不需要担心需要处理新协议和发现算法。您不需要冒险在不确定的 IP/以太网 NVMe 协议上投入培训和教育资金，而如果采用其他的竞争性技术，您势必需要投入此类资金。话虽如此，您应该在采取下一步行动之前，先检查一些现有的基础架构：

- 您可以在现有的 IBM b-type 或 Brocade Fibre Channel 基础架构上部署 NVMe over Fibre Channel，前提是这些基础架构是比较新的版本（FOS 8.1.0 或以上版本）。与硬件供应商一起检查可互操作性。

- 光纤通道结构必须是第五代（16 Gbps）甚至更先进的第六代（32 Gbps）。当然，第七代 (64 Gbps) 也支持 FC-NVMe。
- 使用 NVMe over Fibre Channel 的服务器需要第六代 (32 Gbps) 主机总线适配器 (HBA)；第六代 HBA 也兼容第五代结构。
- 您需要一款支持 FC-NVMe 帧类型的存储设备。它可以是一个基于 FC-NVMe 的阵列，也可以由搭载 FC-NVMe HBA（在目标模式下运行）的服务器扮演存储目标的角色，后者有助于您尽早熟悉该技术。

这些要求都是合理的要求。比如，如果您的服务器正运行性能敏感型应用，而您依然在使用第四代（8 Gbps）光纤通道，那么不论您正在实施哪种 NVMe over Fibre Channel，您都必须立刻升级光纤通道。您获得了实现这一目标的最佳机会。

考虑您的采用战略

除开 NVMe 宣传大师们的大胆预测，很少有存储和网络社区的从业者会质疑以下观点：向 NVMe over Fabrics 迁移是一个循序渐进的过程，整个过程将持续多年。遗憾的是，这种缓慢的转变是一个痛苦的过程，因为眼睁睁将原有的光纤通道网络放在一边，转而实施基于以太网的新 NVMe 网络是一种非常痛苦的体验。在您打算购买更多存储时，您会购买哪种存储？构建新应用时，您会将新应用与哪个环境互联？如果您选择以太网，您会选择以下哪种：iWARP、RoCEv2 或 NVME over TCP？以上哪个选项都没有明显的优势；在存储用途方面，没有一个选项有辉煌的采用记录。但是，通过采用双协议光纤通道结构（运行并发 FCP 和 FC-NVMe 流量），您可以消除或简化上述问题。光纤通道得到了广泛采用，赢得了存储供应商和其他技术提供商的大力支持，这使得 NVMe over Fibre Channel 成为了一个风险相对更低的主张。

保护高价值资产

尽管 NVMe over Fibre Channel 是一项新技术，用于大多数以存储为导向的用途（与“工作内存”不同），其目标是应用该技术处理现有存储资产。随着大数据分析、数据挖掘、数据湖、人工智能、机器学习和深度学习这类概念赢得了越来越多人的关注，每个人数据资产的价值也随之迅速攀升。这使得一流的高性能存储解决方案变得前所未有的重要，与此同时，这也凸显了将风险最小化的重要性。

如果应用仅使用数据资产副本（而不修改或更新主副本），那么应用就是将该数据资产当作工作内存。相反，如果应用将维护主数据副本，那么维持数据的完整性和可用性就变得非常关键了。

大多数情况下，一旦涉及现有的数据资产，那么您就会将目光投向“现有”（而非“全新”）场景。在实验室中完成验证后，您应该循序渐进地将原有架构（通常构建于 SCSI 基础架构之上）迁移到基于 NVMe 的架构，并且提供回滚架构变更的选项。理想的采用战略包含一个支持此类模式的流程和基础架构。

支持马拉松式转变

NVMe 的速度和全闪存阵列的迅速崛起成为了热门话题，那是不是意味着到下周末全世界都会转而采用 NVMe 存储呢？答案是恐怕不会。考虑一下所有宣称的 IPv6 优势以及广泛采用该技术所需的时间。大多数基础架构早在多年前就已准备好采用 IPv6，并且也获得了全面的软硬件支持，但是整个转变依然进展缓慢。即使 NVMe 采用速度大幅加快，整个转变仍会持续数年。

实际上，有些企业或部门在他们只对工作内存感兴趣，没有维护高价值数据资产的迫切需求时，会迅速采用 NVMe。通常，他们将采用直接连接的 NVMe 产品来处理这种情况，在这个阶段他们不需要结构。假如其他企业没有加速工作内存的迫切需求，他们可能先采用 NVMe 用于存储用例，当然，他们也会结合使用多种技术。问题在于，用途的不同，技术的采用速度也不同。

充分利用双协议 FCP 和 FC-NVMe

IT 人员负责企业数据和生产力，因此，他们高度关注风险的降低（用行话说就是“去风险化”）。

IT 部门计划在生产环境内部署基于 NVMe 的阵列时，他们有两个主要选项。他们可以构建一些全新的 NVMe 基础架构（参见图 3-1），或者他们也可以利用现有的基础架构（参见图 3-2）。如果他们构建全新的独立基础架构，自此之后的每一次采购都将是一场赌博，因为他们必须决定哪些基础架构将访问阵列。正因为此，双协议并发对于 NVMe over Fibre Channel 至关重要。

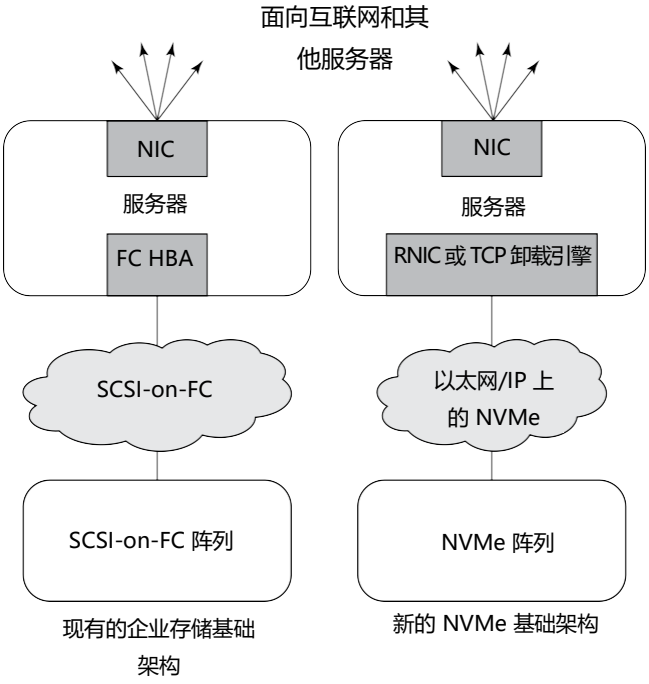


图 3-1： 新的独立基础架构（不推荐）。

通过利用“已知量”，比如 Fibre Channel SAN（存储区域网络），企业能够轻松支持双协议，并消除一个不可回避的问题（“转变将持续多久”）所带来的任何风险。

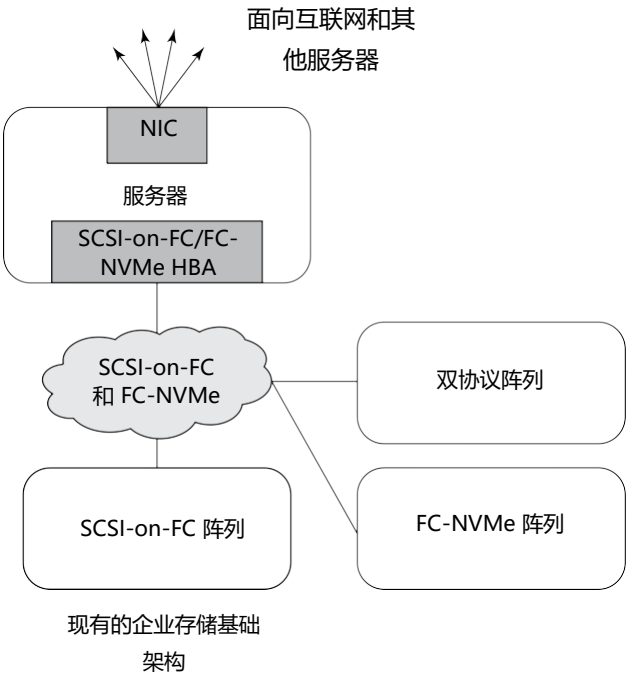


图 3-2： 双协议基础架构（推荐）。

双协议方法合情合理。在光纤通道上同时支持多个协议早已有成功先例。自 2001 年起，在同一光纤通道结构上同时支持 FCP 和 FICON 流量已经成为惯例（参见边栏“FICON 和 FCP”）。

您还需要考虑其他重要事项：

- **循序渐进的迁移：**通常，一个应用会使用多个存储卷，它对这些卷可能有不同的要求。在双协议环境中，单个卷可以视情况进行迁移。不能承受风险的高价值资产依然保存在可信的基础架构上。价值更低的延迟敏感型卷则可以迁移到最热门的新目标上。您可以轻松撤销变更。在双协议环境中，您不需要执行颠覆性的硬件或电缆变更，就可以实现并管理这些变更。
- **双协议发布：**您可以在可信的传统阵列上维护高价值数据资产的主副本，并定期将它们“发布”到最新的高速 FC-NVMe Channel 阵列上，允许其他应用将这些数据当作“工作内存”使用。您也可以在现有的基础架构上做到这一点。

光纤通道允许 IT 管理员使用熟悉的元素，比如分区和名称服务。此外，双协议并发为您提供了千载难逢的机会。

FICON 和 FCP

FCP 是在开放式系统中使用的 SCSI-on-FC 协议，比如 Windows 和 Linux 系统。FICON 是大型机 (z/OS) 存储协议的光纤通道版本。很多使用大型机的企业都经常需要“发布”大型机数据资产，以便开放式系统能够使用这些数据资产。而在其他时候，比如在并购企业后整合数据资产时，企业可能需要将 FCP 数据资产融入 FICON 环境中。

为了传输这些资产，企业为专门的迁移服务器配置了连接双协议 SAN 的双协议（FCP 和 FICON）HBA，双协议 SAN 将连接 FCP 存储阵列和 FICON 存储阵列。

事实上，FCP/FICON 双协议 SAN 的目的并不是长期从一个协议转变为另一个协议，而是在两个协议之间提供持久的“桥梁”。因为大型机和开放式系统的结构不同，所以业内并没有将应用从一个协议迁移到另一个协议的概念，普通的应用服务器的配置也不支持同时使用两个协议。

并发双协议 FCP/FC-NVMe 则截然不同，因为两个协议专为开放式系统量身定制，借此，您可以轻松制定计划，以不中断业务的方式循序渐进地将应用从一个协议迁移到另一个协议。有些应用能够在更长的时间范围里在双协议模式下运行，利用一个协议消费（读取）资产，然后用另一个协议发布（写入）资产，这会让这些应用受益匪浅。您也可能选择相对快速地（在几天或者几周内）转变其他应用，即，那些在双协议模式下花费时间相对很少的应用。不论采用哪种方式，通过利用双协议 FC，您都能在将 NVMe 部署到生产环境时，大幅提高灵活性。

分区和名称服务

网络管理员利用*分区*提高安全性。目前，您可以采用两种分区方法：

- **端口分区**基于节点连接的光纤通道设备的特定端口，限制访问。
- **名称分区**基于设备的全球通用名称 (WWN) 限制访问。

两种方法都限制了设备访问不该访问的网络区域。哪种方法更适合您？这取决于您的使用情况。名称分区往往能帮助您减少维护工作，除非一个端口连接了一系列不同的设备。

您可能会碰到引用*硬分区*和*软分区*。现代化 FC 结构使用硬分区，在硬分区内，强大的硅基逻辑会拦截节点之间不允许传输的流量。相反，早期的产品使用软件来实现软分区。软件无法拦截流量，只能隐藏信息。这就好比您只能掩盖您的地址，而不能锁门。这也就难怪软分区被淘汰了。

重点在于，FCP 和 NVMe over Fibre Channel 都能使用 FC 分区。光纤通道的分区服务在结构中实施，这种方法与同类竞争性技术所用的方法截然不同。因此，结果更加可预测，更好管理，并且减少了安全漏洞的机会。

另一方面，名称服务将晦涩的计算机和设备地址翻译成人性化的名称。这些名称类似于 DNS，但是常驻网络，从而大大简化了可互操作性和管理。FCP 一直如此，现在，这一点同样适用于 FC-NVMe。

发现与 NVMe over Fibre Channel

NVMe over Fabrics 的规范描述了一个发现机制，但是在具体的实施上还欠缺很多细节。这给非 FC 结构留下了一条巨大的可互操作性鸿沟，并且这条鸿沟的解决速度可能会像解决以往的可互操作性挑战一样慢，比如基于优先级的流控制 (PFC) 和数据中心桥接交换 (DCBX)。

光纤通道能够交付双协议 FCP/FC-NVMe 结构，这相当于提供了一条方向清晰的前进道路来提高可互操作性。HBA 供应商正在创造驱动因素将 FCP 用于设备发现，然后检查这些设备是否支持 FC-NVMe 流量。提供 SCSI-over-FC 阵列的企业存储供应商也有动力支持这种两步法。新兴的 NVMe 阵列供应商如果对 FC 市场感兴趣，他们可以利用从 SCSI 映射到 NVMe 的 NVM Express 标准，为了吸引现有的 FC 客户，他们最有可能遵循这一模式。

熟悉 NVMe over Fibre Channel

终有一天，NVMe over Fibre Channel 也将成为一项被人们所熟知的技术。但是至少目前，它对于人们来说更像一项从未谋面的技术。您不确定该技术将如何应对特定的刺激，它需要多少关注，或者它是否会突然在紧张的情境下反应过度。

别担心。有关这项激动人心的技术，您已经尽你所能了解了所有能找到的知识，现在，是时候卷起袖子放手干了。留出时间来熟悉 NVMe over Fibre Channel 的细节。查阅供应商的可互操作性矩阵，确定您的具体设置如何运转。然后，跳出测试环境，思考 NVMe over Fibre Channel 如何配合您的生产系统。

在实验室试用

在第一次试用 NVMe over Fibre Channel 测试环境时，您可能想像弗兰肯斯坦博士一样大喊：“它是活的！”请克制住这种冲动，因为这可能让其他人觉得您的成功让您惊喜不已。相反，请平静地点点头说：“是的，这就是我想说的，”然后给自己开一瓶新的红牛。

不确定从何处着手？通常，IT 部门在成立 NVMe over Fibre Channel 测试实验室时，可以采取以下步骤：

- 设置一个服务器，并搭载内部驱动器、单一交换机和支持 FC-NVMe 的单一阵列。
- 将服务器与实验室 IP 网络互联，以便访问其他实验室服务器。

- 探索您已看到很多次的 HBA 配置和存储阵列选项。
- 配置一个卷，以便使用 NVMe 阵列。具体步骤取决于您的 NVMe 阵列管理工具，但是通常，该步骤与配置 LUN 类似。NVMe 中的命名空间 ID (NSID) 就相当于 SCSI 内的 LUN。
- 将文件从内部驱动器复制到 NVMe 卷，并反复执行这一操作。
- 运行首选的性能测试应用（比如 Iometer），以便对您的 NVMe 卷进行基准评估。将它与您的内部卷进行比较。
- 不断重复，直到消除您的疑虑。

将 LUN 迁移到命名空间

将海量数据从一项存储技术迁移到另一项存储技术绝非易事。尤其当您对这个流程不完全清楚时，这项工作更是难上加难。从小处入手。将一两个卷从 SCSI 迁移到 NVMe（从 LUN 到命名空间 ID），从而建立起对该流程的信心。

在实验室里，您还可以在 NVMe 卷（命名空间）上运行一些应用，形成一种“肌肉记忆”，确保您不会忘记任何步骤。这时，您应该隐约而兴奋地感觉到您已经了解了整个工作原理。

下一步就是提高这种愉悦度。您已经完成了基础工作，下一步，请打开所有存储管理应用、SAN 管理应用和分析应用，比如 IBM Network Advisor、Brocade SAN Health 和 IO Insight。确保您已经对这些工具进行了适当的升级，它们都支持 FC-NVMe。此外，您还应该确保您知道 NVMe over Fibre Channel 给这些应用所带来的改变。最后，思考一下在线将 FC-NVMe 融入生产环境时，您需要其中哪些工具和功能。

在实施任何重大的生产变更时，您应该考虑衡量“基线”性能。请确保第二章探讨的主机 CPU 利用率和存储指标也涵盖在内。谨记，切换到生产环境后几个小时，您可能就会收到支持请求，因为一些传统的应用出现了问题。这只是常规的人为失误，还是真正的 hiccup 错误？如果是 hiccup，是不是切换到生产环境导致的？这时，您可以喊出童子军的口号：时刻准备着！到目前为止，您应该已经有了足够的信息来确认支持请求是否与最近的变更有关。您早前确定的基线能助您一臂之力。

即使没有支持人员的电话指导，您也会想要知道通过从 SCSI 迁移到 NVMe 性能提高了多少，因为这一信息有助于您评估其他行动，并确定其他行动的优先级。此外，当用户因应用的速度而欣喜若狂，向您发送电子邮件时，您能够准确地告知他们系统的运行速度到底提高了多少。好吧，我们知道不可能出现这种情况，但是至少您可以赢得上司的一句“做得好”的表扬。如果您不知道初始性能，您就不太好高举双手，宣传 NVMe over Fibre Channel 的超凡状态了。

一切都很好，您现在成了 NVMe over Fibre Channel 绝地武士，但是如果您休假了或者晚上出问题了，怎么办呢？除非您想在阿尔卑斯山滑雪时接到工作电话，否则请确保在上线之前，您的同事也跟上了您的脚步，并且所需文档已准备就绪。


提示

查看第五章，了解 FC-NVMe 的最新性能增强包。

迁移到生产环境

抓紧了，要上线了！将测试系统迁移到生产环境是一段激动人心的经历，尽管过程让您揪心，并且您可能需要通宵加班。鉴于您已经信心满满，也有了充分的了解，那么是时候开始挑选哪些应用和卷最适合上线了，并对这些应用和卷进行优先级排序。

就像您在实验室一样，请确保所有管理工具都进行了妥善的升级，能够支持 FC-NVMe。您总不希望在踏上跨国之旅后，发现您忘带了地图，油箱也空了，前胎也憋了。您肯定急于分享您的劳动果实，但是请不要在这最后（往往也是最重要的）一步上侥幸偷懒。

最后，请安排在一个不会干扰客户的时间进行迁移（他们会在您告知需要重新引导时表现得毫无兴趣），确保企业内的所有人都了解即将到来的迁移。所有人（尤其是您）都不应该感到意外，你们都应该享受在 NVMe over Fibre Channel 这条高速公路上的旅程。

本章提要

» 远程直接内存访问的优缺点

» InfiniBand

» iWARP

» RoCEv2

» 评估基于以太网的 NVMe

第 4 章

比较 NVMe over Fibre Channel 与其同类产品

在大多数情况下，有替代方案总归是好事。当您在选择主卧铺的地毯颜色或者高峰时期的回家路线时，有替代方案绝对是好事。NVMe over Fabrics 也提供替代方案，尽管某些方案可能不是您中意的。您可以在 iWARP、RoCEv2 或 InfiniBand 等结构上运行 NVMe，或者您可以直接运行 NVMe over TCP。本章将介绍每个同类产品的优缺点，并探讨性能考量因素，比如网速、架构、虚拟化，以及支持的专用功能。当然，面对风险，性能毫无意义，因此，本章也会评估可预测性和潜在的干扰。

远程直接内存访问的优缺点

RDMA 是远程直接内存访问的简称。该协议问世于几年前，初衷是在“紧密耦合”的服务器环境中使用，尤其是高性能计算 (HPC) 类别的服务器环境。如果人类用 RDMA 交流，那么我们不再需要语言或肢体语言，思想和情感将直接在大脑之间共享，进而显著提高交流速度，消除任何产生误解的机会。

幸运的是，人脑不是电脑，我们可以保留自己的想法。但是，对于集群型服务器应用来说，RDMA 是一种共享动态信息的绝佳方式。一个服务器将一部分内存的“所有权”转交给一个远程服务器。对于很多多服务器应用，尤其是那些需要动态变更数据的应用来说，这种方法能够提供显著的性能优势。

宣称零副本

20 世纪 80 年代 TCP 堆栈问世时，市场上已经有各种各样的网络技术。因此，TCP 堆栈旨在与任何可用的网络技术协同运行，不论是令牌环还是电话线。包含清洁网络层对于可互操作性意义重大，要做到这一点，您可以使用中间缓冲，这也使得缓冲副本变得非常普遍。但是，随着速度的加快，大多数缓冲副本都被优化掉了，但该实践会破坏向后兼容性的情形除外。

在 20 世纪 90 年代中期，一个良好的网络堆栈就能保证单一副本的效率。网络适配器接收帧并（利用 DMA）将它们写入与网络堆栈有关的 DRAM 缓冲。（不可避免的 DMA 步骤并非 DRAM 到 DRAM 的复制，因此，该步骤不计入在内。）首先，网络堆栈处理帧，然后将“有效负载”复制到高级应用期望的内存位置。在一段时间里，单一副本架构似乎得到了全面优化。

但是，当 FC 进入“产品化”进程时，游戏规则开始改变。光纤通道宣传的主要优势是速度，因此，光纤通道迫切需要提高速度。芯片技术能允许更高的复杂性，限制 IP 堆栈的向后兼容性挑战却无法限制光纤通道/SCSI 堆栈。FC 专注于一个“应用”（存储），其层结构比 TCP/IP/以太网更简单。由于所有这些原因，光纤通道得到了更多的激励，能够更好地实施网络适配器/驱动器/堆栈结构来消除单一副本。实际情况正是如此。过去 20 年间，光纤通道一直在默默地交付“零副本”。

NVM Express (https://nvmexpress.org) 主编了一份白皮书，旨在描述两类面向 NVMe 的结构传输方式：使用 RDMA 的 NVMe over Fabrics 和使用光纤通道的 NVMe over Fabrics。虽然 FC 被明确视为 NVMe 结构，但是有些 RDMA 支持者会宣称，因为 NVMe over Fibre Channel 不使用 RDMA，所以它在某种程度上并非 NVMe 结构，尽管事实是 NVMe over Fibre Channel 不需要 RDMA，我们将在本章后面详细探讨这一点。FC 使用本地直接布局功能，同时提供双协议支持，让您能够以低风险的方式从 SCSI 迁移到 NVMe。如果您对这些宣言有任何保留意见，请继续往下看：我们将一一消除您的疑惑。

InfiniBand

InfiniBand (IB) 的面世比以太网或光纤通道要晚。IB 聚焦服务器集群通信，致力于在本地交付 RDMA。相比大范围采用，IB 更重视速度。您必须有特殊的适配器和交换机才能使用 IB，之所以 IB 没有得到广泛认可，也没有与合作伙伴产品实现广泛兼容（专用 HPC 应用除外），这也是原因之一。事实上，在写这篇文章时只有一家 IB 芯片供应商提供 InfiniBand 产品，这一现实会挫败新采用者的积极性，使得人们质疑切换至 InfiniBand 协议的成本。这家 IB 芯片供应商似乎也意识到了这一点，因此，该公司大力开展 NVMe over Fabrics 营销工作，以推广 RoCE。

iWARP

iWARP 问世已近十年，尽管目前也没有得到广泛采用。iWARP 全称为“互联网广域 RDMA 协议”，它是 2007 年五个 RFC 中描述的 IETF（互联网工程任务组）标准，最近一次更新在 2014 年，新增了三个 RFC。iWARP 运行于 TCP 之上，它被归类为可靠的流媒体传输协议，因为其中包括不同的技术，以确保发送者收到了发出的每一个字节。但是，iWARP 的 TCP 基础并不太适用于存储，因为 TCP 通常需要慢慢提升传输速度。鉴于很多存储应用的流量模式都包含所谓的“突发性大象流”，TCP 的慢启动行为会衍生出延迟挑战，并降低 IOPS 指标。

TCP 经过了多次扩展，新版的 TCP 堆栈提供可配置（有时甚至可协商）的功能，而老版本则没有这类功能。早期 TCP 采用“慢启动”机制，以缓慢启动传输，避免网络缓冲区出现溢流。如果协议超时或者接收者 ACK 消息显示未收到某些传输的数据，传统的 TCP 会重新传输数据，并撤回一些传输带宽（因为它“使得 TCP 窗口崩溃了”）。Data Center TCP (DCTCP) 这类新版的 TCP 提供的功能可在数据中心环境下更好地运行，但是这些功能与 WAN 使用不兼容。正因为此，如果数据中心的架构师和实施人员想利用 TCP 处理高性能用例，他们就会面临挑战。这时，他们有三个选项：

- 选择一个复杂的 TCP 堆栈，根据不同的用例对该堆栈进行不同的配置，并跨越所有操作系统映像强制实施该复杂的堆栈
- 选择两个或更多个 TCP 堆栈，并管理哪个映像获得哪个 TCP 堆栈
- 部署一些获得两个（或多个）TCP 堆栈的 OS/管理程序映像，在内部（可能是按 IP 地址）映射至预期的使用场景

以上每个选项都存在问题，都会增加复杂性，给网络管理员带来额外的负担。

性能会衍生出最后一个缺陷。TCP 的初衷是跨越各种网络运行，其中当然包括广域网（如 iWARP 这一缩写中所述）。但是为了有效地跨越各种网络运行，TCP 会尽可能减少因发送者发送消息过快导致的消息丢失，这通常会导致“减速”。正是这三个原因导致 iWARP 没有被网络社区广泛采用。

技术内容

作为一个可靠的流媒体协议，独立的 TCP 从来无意保证数据包或帧的一致性。这是因为 TCP 发送的不是一系列数据包，而是一连串单独的字节，这也就无法保证 8 字节的数据包中的命令能够得到及时处理，因为首先需要解码整个信息流。TCP 是一个基于软件的复杂堆栈，因此，一致性问题会阻碍 iWARP 的硬件处理。为了解决一致性问题，其中一个 iWARP RFC (RFC 5044) 构建了一个修复方案（“Marker PDU Aligned Framing”，也即 iWARP-MPA），旨在以增加堆栈复杂性为代价，确保数据包的一致性。

这不免让人想起了被 NVMe 所取代的麻烦的 SCSI 堆栈。

你好，Rocky!

RoCEv2 是 RDMA over Converged Ethernet（版本 2）的简称。它的发音是“Rocky vee two”，如果您在服务器房里把这个词读错了，整个房间的人都会哈哈大笑。这是一个奇怪的标准，因为它是由 InfiniBand Trade Association 而非 IETF 或 IEEE 制定的，而大多数 IP 和以太网标准都是由 IETF 或 IEEE 负责制定和维护。

RoCEv2 全称为“RDMA over Converged Ethernet”加上“版本 2”，它运行于 UDP 之上，因此，它不再与以太网直接相连。但是，为了获得性能和可靠性，RoCEv2 建议使用 Converged Ethernet，这是一个过去的术语，指的是一种无耗以太网网络。现在，无耗以太网被正式命名为数据中心桥接，其中包括很多相互依赖的功能，比如基于优先级的流控制、增强的传输选择和数据中心桥接功能交换。

DCB 是一项持续工作，旨在利用 20 世纪 90 年代中期开发的光纤通道中的功能，增强以太网。尽管这个目标值得称赞，但是实际部署的 DCB 的可互操作性依然很低，再加上以太网/IP 的宽容特性，这使得问题变得更加复杂。您可以在很长一段时间内都无法轻易发现配置错误的 DCB 网络，平时它会像正常的以太网网络一样运行，但是一旦出现流量高峰，它就会随机丢失数据包。

如需查看基于以太网的 NVMe 选项与 NVMe over Fibre Channel 两者之间的差别，请查看表 4-1。

表 4-1 基于以太网与光纤通道

基于以太网的 NVMe 选项	NVMe over Fibre Channel
制定新的结构协议标准	基于 T11 标准化光纤通道结构协议构建
标准组织负责解决将 I/O 命令、状态和数据扩展至数据中心的挑战	针对 SCSI 制定 FCP 协议后，光纤通道负责解决这些问题

（未完待续）

表 4-1 *(续)*

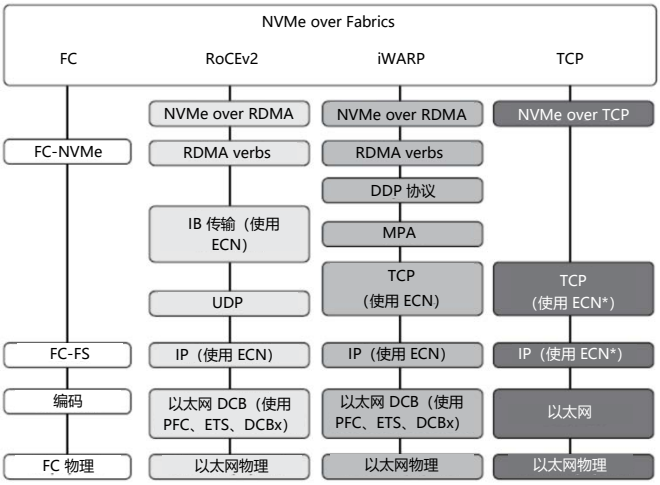
基于以太网的 NVMe 选项	NVMe over Fibre Channel
传输选项： iWARP、RoCEv2 和（最近的）TCP	光纤通道负责传输；运行于现有的 ASIC 之上
iWARP 和 RoCEv2 使用 RDMA（TCP 不使用 TDMA）	不需要 RDMA；利用 FCP
如果启用 RDMA，则网络配置很复杂	光纤通道得到了充分理解
新的 I/O 协议，新的传输	新的 I/O 协议，现有的可靠传输
如果一起使用 RDMA 与 RNIC，则延迟率低	使用 FCP 时同样实现零副本性能
发现和分区服务依然处于建议阶段	利用久经验证的结构服务

评估基于以太网的 NVMe

为存储网络的物理层选择以太网难度很大，原因有多个。首先，业内主要关注的是以下四个基于以太网的 NVMe 协议：iWARP、RoCEv2、NVMe over FCoE 和 NVMe over TCP。为了在使用 iWARP 或 RoCEv2 时实现低延迟，您需要安装支持 RDMA 的 NIC (RNIC)。Facebook 这类大型运营商会选择以商品为导向的 NVMe over TCP（相当于如今的 iSCSI），即使它速度更慢。当然，您购买的 NVMe 阵列必须能够支持您的结构选项，但是谁又知道哪个协议会成为最后的赢家。结果就是，在尘埃落地之前，不论选择哪个基于以太网的 NVMe 协议都存在风险。

其次，与 NVM Express 组织在其相关 NVMe over Fabrics 白皮书上所提的建议相反，以太网流控制并未使用在光纤通道、PCI Express 和 InfiniBand 传输中发现的可靠且基于信用的流控制机制。

另外，不论您是选择 iWARP 还是 RoCEv2，您都在选择一个多层网络，随之而来的就是传输 NVMe 时堆栈复杂性增加（参见图 4-1）。以太网支持者大肆宣传巨型帧这类优势，尽管 Demartek 等机构建议在使用 RoCEv2 时禁用巨型帧。有些数据中心正在使用 VXLAN，这会增加额外的以太网和 IP 报头，并且需要设置“最大协议数据单元”(MaxPDU) 来额外管理每个网络端口。MaxPDU 会影响 IP 分片，这反过来又会分别影响 IPv4 和 IPv6。为什么需要这些层？一部分是因为遗留问题，另一部分是因为以太网/IP 是专为互联网规模而非光纤通道的数据中心规模设计的。NVMe 的主要优势源自其简化的架构，这种情况下您还选择复杂的多层传输就会非常奇怪。



*NVMe over TCP 不需要 ECN，但是这是一种普遍用于避免 NVMe over TCP 丢包的方法。

图 4-1： 各个 NVMe 结构堆栈的相对复杂性。

商品还是优质产品？

以太网是一项低成本且会尽力运行的技术。因此，它成为了市场上的赢家。只需轻松部署即可实现常见用途，它支持各种上层协议和应用。以太网的即插即用式特点专为广泛采用而设计，业内数以万计的资深技术人员都知道如何管理主流以太网配置。以太网拥有强大而又过分简单的机制，比如生成树协议，该协议能够关闭链路，保证没有任何回路能引起广播或组播风暴等问题。否则这些问题就可能变得非常普遍，因为广播是地址学习的普遍组成部分。

遗憾的是，基于树的拓扑结构并不太适合用于处理当今的数据中心流量。因此，企业倾向于在服务器机架上使用 IP 路由器。以太网部署很容易，很大程度上是因为其主要客户 TCP/IP 太有弹性、太宽容，这牺牲了本就中等的性能。第二层以太网并未大范围扩展，但是连接 IP 后，它能够扩展至互联网。以太网和 IP 随处都可以买到。eBay 和亚马逊上就有出售一些有趣的产品，这也使得它们成为了一个极具竞争力的销售平台。光纤通道网络产品主要由存储供应商提供，要引起供应商大打价格战并不是那么容易的事情。

智慧地购物

另一方面，大多数存储供应商都投入了大量时间，用他们销售的网络产品测试他们的阵列。对于集成至 ASIC 和管理软件的所有增强型分析和可视性功能，他们都非常熟悉。存储供应商非常了解常驻网络的功能，比如光纤通道名称服务和光纤通道分区，包括由目标驱动的分区，这些功能尚未针对基于以太网的 NVMe 结构进行定义。这些供应商非常擅长处理与以下方面相关的支持问题：服务器、HBA、存储阵列和网络之间久经验证的相互影响。

提示

如果您选择购买价格最低的以太网/IP 网络，那么您需要考虑一下当您因为一些奇怪的问题而打电话给存储提供商时，您会获得怎样的支持。从何处着手？如何检测网络以着手解决问题？除了企业存储应该部署在专用网络上这一众所周知的建议外，IP 存储经常与共享网络互联。有鉴于此，请考虑调查您现有的以太网/IP 网络，评估您是否能够在此类网络上支持存储 SLA。

以太网和 IP 的大获成功是一把双刃剑。当您需要一个更专门的优质网络时，很多让以太网和 IP 如此普及和商品化的因素却变成了麻烦。在互联网、校园、家庭和移动设备这类光纤通道没有意义的地方，以太网和 IP 这两个协议套件无疑是首要选择。甚至以太网和 IP 也非常适合在数据中心内使用，因为它们能满足多协议尽力通信需求。但是，数据中心内的企业存储是一个要求更苛刻的用例。这也是为什么“足够好”还远远不够，投资宝贵的资产才有意义。

本章提要

- **知道可以期待的性能改进成果**
- **选择最短的捷径实施 NVMe over Fibre Channel**
- **考虑 SAN 设计时兼顾 FC-NVMe**
- **什么是 ANA？为何 ANA 如此重要？**
- **探索应用用例**
- **知道并非所有结构都是一样的**

第 5 章

利用 NVMe over Fibre Channel 改进性能

现在您准备试着采用 NVMe over Fibre Channel 了，那么预测相比 SCSI/FCP 您将得到多大的性能改进成果，原因为何，这是一个大问题。为了帮助您解答这个问题，本章将探讨您可以期待的性能改进成果。NVMe 旨在充分利用特点类似于内存的闪存存储。因此，它是一个比 SCSI 更高效的协议。然后，本章将介绍如何将这种方法应用于端到端的解决方案，包括服务器上运行的应用、光纤通道 SAN 和存储阵列等等。

了解 FC-NVMe 如何改进性能

图 5-1 展示了一条端到端的存储链，它从主机上的应用延伸到了存储阵列上的存储介质。NVMe over Fibre Channel 在以下领域性能更佳：

- **主机端**：相比 SCSI/FCP，FC-NVMe 在服务器上表现如何
- **存储阵列前端**：相比 SCSI/FCP，FC-NVMe 在存储阵列目标端口上表现如何
- **存储阵列架构**：相比 SCSI/FCP，存储阵列架构如何处理 NVMe
- **存储阵列后端**：相比 SCSI/FCP，如何通过用 NVMe 连接的 SSD 替代 SAS/SATA 连接的 SSD，来改进性能

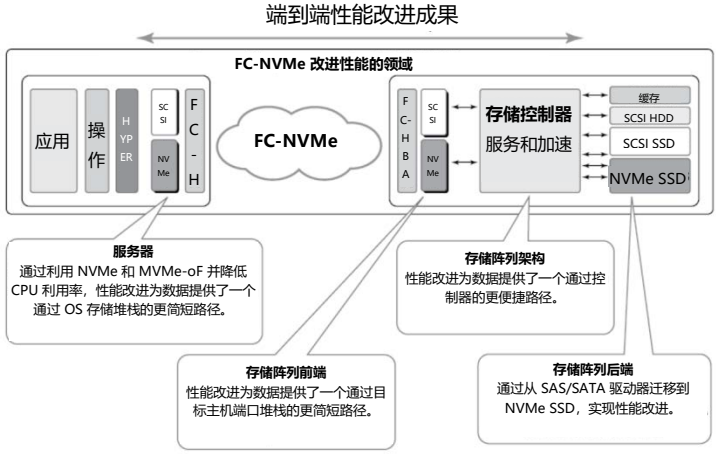


图 5-1：借助 NVMe over Fibre Channel 改进性能的领域。

结构怎么样?

你可能会想，“哇！但是，等等，光纤通道 SAN 怎么样？”切记一点，SAN 为 SCSI (FCP) 和 FC-NVMe 提供同样的支持。光纤通道网络在传输 FCP 和 FC-NVMe 时的性能没有差别。

您会体验同样低延迟、高性能的传输。但是，光纤通道标准增加了增强包，以便在出现低级错误时增强高性能 FC-NVMe（参见边栏）。下面，我们将详细介绍每个领域：

主机端

在主机端，您会发现 CPU 利用率有了大幅提升，因为 NVMe 命令集比 SCSI 更精简，每个核心有多个队列，每个队列有多条命令。同时，它还有单独的异步提交/完成队列，这样，它就形成了一个更精简、更平均的驱动器堆栈，该堆栈使用的资源比 SCSI 少，执行速度却更快。在处理同一工作负载时，FC-NVMe 需要的 CPU 处理时间比 FCP 更短，进而为您的应用提供更多 CPU 周期。随着时间的推移，应用将利用这种更精简、更平均的驱动器堆栈，进一步降低应用延迟。

面向 FC-NVMe 的序列级错误恢复

什么是序列级错误恢复？为什么序列级错误恢复至关重要？T11 组织正在编写 FC-NVMe-2 标准，标准中包括面向 FC-NVMe 的增强包，名为“序列级错误恢复”。目标是在序列层面恢复错误，而不用将错误上报至存储协议层面 (NVMe)。为了恢复错误，NVMe 启动程序/目标适配器对于重发丢失或受损的命令达成了共识。允许传输层恢复丢失或受损的命令有一个好处，那就是错误能够更快地恢复，且几乎不会甚至完全不会影响存储性能。借助存储级内存，NVMe over Fibre Channel 技术将端到端存储延迟缩短到了几十微秒。这种情况下，该技术能够提供比其他任意结构技术都出色的错误恢复功能。

如需获取最专业的资讯，你可以查看 FC-NVMe-2 标准中的其他精彩内容 (https://nvmexpress.org/resources/specifications)：

- 管理命令确定性
- T10 保护信息处理增强包
- 提交队列流控制处理增强包

存储前端

在存储阵列主机端口端（前端），比 SCSI 更精简的 NVMe 协议能让数据更快速地抵达存储控制器。有时，您还会听到一种改进成果：为数据提供一个通过驱动器/协议栈的“更简短路径”。

存储阵列架构

过去，存储阵列控制器通过将 IO 分散至旋转介质并提供存储服务，提高性能。这些服务包括数据保护、加密、压缩和去重，这些服务不会增加任何明显的延迟，因为旋转介质的速度比控制器上运行的存储服务要慢几个量级。

该实践随着 SSD（尤其是 NVMe 连接的 SSD）的改变而改变。相反，从延迟的角度来看，存储服务正在崭露头角。市场上涌现的新阵列架构旨在让阵列控制器远离数据路径，针对延迟要求低的应用，取消选定存储服务。

提示

有很多令人信服的理由驱使我们在现有的阵列上切换至 FC-NVMe。只需要将现有的存储阵列升级到最新的固件级别，您就能使用便利的 FC-NVMe，这是一种直接采用和实施 NVMe over Fibre Channel 的方法。

存储阵列后端

鉴于目前几乎所有市场上的阵列都使用 SAS/SATA 连接的 SSD，因此，您有机会利用 NVMe 连接的 SSD 改进性能，因为 NVMe 连接的 SSD 能够交付更低的延迟和更高的 IOPS。谨记一点，该实践发挥价值的前提条件是总体阵列架构都能通过阵列端到端地交付性能。假如阵列的设计目标是端到端地交付最低的延迟和最高的 IOPS，那么在阵列中采用最新的闪存介质技术（比如 3D-TLC）和存储级内存（SCM）（比如 3D-Xpoint，发音是“three dee cross point”）就很有意义。

PCIe Gen4 和 Gen5 是另一项崭露头角的新技术，它们将在新的存储阵列设计决策中起作用。如需详细了解 PCIe Gen4 以及它在 NVMe 情境下的意义，请参阅边栏“PCIe Gen4”。

PCIe Gen4

2018 年中上市的全新 NVMe SSD 与 PCIe Gen4 兼容。PCIe Gen 4.0 标准规定了 PCIe Gen4 的数据链路速度为 16 Gbps，最高有 16 个链路或线路交付 64 Gbps，这相当于将 PCIe Gen3 的最大速度 32 Gbps 翻了一番。一旦 PCIe Gen4 x86 主板上市，市场将开始推出性能无与伦比的全新 NVMe 存储阵列。PCIe Gen5 标准正在起草中，目标是在 2019 年出台。PCIe Gen5 将通过 16 条线路交付 128 Gbps。

切记

此外，如果重构多线程应用和管理程序，以便充分利用 NVMe 的多队列属性，那么随着时间的推移，您也可能从这一实践中实现改进。

通过升级软件处理 NVMe 支持

市场上第一款交付 FC-NVMe 的存储阵列能让您轻松着手使用 FC-NVMe。您只需简单地升级软件，将存储阵列控制器升级到最新的固件版本，然后您就可以在存储阵列上同时配置 NVMe NSID 和 LUN 了。通过在服务器中使用第六代或（现已上市的）第七代 HBA，您就已经做好了使用 NVMe over Fibre Channel 的准备。

改进性能

鉴于您已经知道为什么要了解端到端的 NVMe over Fibre Channel 解决方案带来的性能改进成果以及这些改进来自哪些领域，这一部分我们将介绍该解决方案带来了多大的改进。

本文撰写之时，市场上这两种产品的供应商都展示并记录了 FC-NVMe 相比 SCSI/FCP 在主机端所改进的延迟、IOPS 和 CPU 利用率。尽管系统和测试结果之间有一定的差异，但是系统和测试结果都显示，在处理同一工作负载时，应用执行速度提高了 30-50%，IOPS 提升了 25-50%，使用的 CPU 资源减少了 30-50%。系统和测试结果展示的都是事务性数据库的一个普通 OLTP 工作负载概况。

考虑针对 NVMe 量身设计的应用

想象一下，有一个程序能够将一个数学问题分解成多个并行的流数据请求（可能最高为 32 个请求），这些请求将被返回至芯片（可能是一个 NVIDIA 芯片）加以处理，并且该程序有多个 GPU。从功能的角度来看，GPU 就是大型浮点引擎。然后，如果应用能够整合 32 个数据流的输出，那么这会对财务分析、威胁分析或渲染的周期时间产生何种影响呢？您将获得无限的可能性。

您可能会想："哇！这太棒了。通过在支持 FC-NVMe 的阵列上简单地升级软件，您就能将应用执行速度提高 30-50%。"那么，问题来了。如果光是存储链速度更快就能让您满意，那么您将错过 NVMe over Fibre Channel 带来的最大机遇。

第 2 章介绍了增强型队列，其中 NVMe 支持 64k 的队列，每个队列有 64k 的命令。如何构建增强型队列呢？答案就是设计多线程现代化应用，以便充分利用 SSD 和 NVMe 协议，来使用多个并发线程执行 IO，支持并行队列。

当然，我们需要一些时间来重新设计所有应用，以便以最优方式利用 FC-NVMe 和 SSD。这种改变在哪类应用或哪一层非常明显，或者哪类应用或哪一层有可能很快实现这种改进？答案就是负责在数据中心内虚拟化服务器硬件的管理程序层。通过增加并行存储 IO 线程，您的存储性能可能比虚拟机高一个量级。

考虑 SAN 设计时兼顾 FC-NVMe

您可能还会思考一个问题：NVMe over Fibre Channel 如何影响 SAN 的设计？从 SAN 设计的角度来看，您需要关注的领域在某种程度上都是相关的。

通过以同样的方式在数据中心内转而采用全闪存阵列，您可以提高全闪存阵列存储端口的 IO 密度，将每个端口的 IOPS 提高几个量级。主机与目标端口的比值也将提高。FC-NVMe 将进一步提升这种改进。随着主机端口与存储目标比值的提高，以及每个存储端口 IOPS 的提升，超额订阅主机的风险将随之增加，这类主机存在瓶颈行为，并且会给其他高性能应用带来不利影响。

提示

*瓶颈设备*指的是无法及时向交换机返回缓冲区信用阀值的主机或存储阵列。这会导致帧被退回结构，进而引起结构拥塞。在结构内，多个数据流共享 ISL 以及 ISL 上的 VC。但是，只要数据流使用的是同一链路上的同一个 VC，那么这些数据流就能使用信用阀值在 ISL 或链路上发送流量或数据包。因此，瓶颈设备可能会放缓信用阈值的返回，阻碍同一链路上数据流的流畅传输。

为了降低瓶颈设备的影响，IBM b-type Slow Drain Device Quarantine (SDDQ) 功能可支持 MAPS 自动识别并隔离瓶颈设备，因为该功能可将流量移动至结构内的低优先级，避免对正常的数据流产生负面影响。在实施 NVMe over Fabrics 时，启用 SDDQ（Fabric Vision 的功能之一）至关重要。

还有一个步骤也很重要：评估交换机端口与 ISL（扇入）的比值，验证是否有充足的带宽来应对流量高峰。借助 FC-NVMe，我们能够轻松将边界推向更高的高度。因此，FC-NVMe 存储阵列的足迹增加时，您可能需要在现有 SAN 内增加交换机之间的 ISL。

了解为何监控至关重要

在管理 SAN 这类高性能基础架构时，监控就是基石。意识到监控的重要性后，您可能已经使用了部分或整个 IBM b-type Fabric Vision 工具套件。在 SAN 内添加 NVMe over Fibre Channel 后，SAN 的监控变得更加重要了，因为通过监控，您可以提前发现问题，避免问题影响应用性能。此外，监控还能帮助您在出现问题时，排查故障，找到问题的根源和解决问题的途径。

启用 MAPS 是底线。通过用 MAPS 来辅助第六代平台上的 IO Insight 功能，您可以获得 Flow Vision 内置的设备输入/输出 (I/O) 延迟和性能检测。在最新的 IBM b-type Gen 6 产品中，IO Insight 功能包括针对存储 I/O 健康状况和性能的 FC-NVMe 协议级别的非侵入性实时监控和警报功能。这些额外的可视性能帮助您深入洞悉可能出现的问题，并帮助您维持服务级别。

配合使用分区

分区能够应用于访问 FC-NVMe 目标，并以同样的方式应用于访问 SCSI/FCP 目标。实施 NVMe over Fibre Channel 能够改变您对 NVMe Controller 和 NSID 访问进行分区的方式。原因在于，一些存储阵列实施了 NVMe Controller 目标端口作为物理目标端口 (WWPN) 背后的逻辑接口（子 WWPN）。

对等分区

对等分区允许“主”设备与分区内的其他设备通信。主设备负责管理对等分区。分区内的其他“非主”设备只能与主设备通信；它们不能相互通信。

在对等分区内，允许主设备与非主设备进行通信，但是不允许非主设备与非主设备进行通信。这种方法能够建立分区连接，进而以简单的方式提供高效的单一启动程序分区，并降低分区数据库内存的使用，就像一对多分区一样。通常，对等分区有一个主设备和一个或多个非主设备，但是对等分区可以配置多个主设备。对等分区与传统分区并不相互排斥；同一个分区配置和结构可以同时存在多个分区风格。

提示

您可能已经很熟悉结构内的 NPIV 登录，大体上，分区的工作原理与 NPIV 登录一样。因此，针对为逻辑接口端口配置了 NSID 的主机，您需要进行分区。这些端口面向的是 NVMe 控制器。

什么是 ANA？为何 ANA 如此重要？

FC-NVMe 提供的多路径 IO 支持是一个令人困惑的主题。但是关键在于，对称型多路径是 NVMe 规范的一部分，它也适用于 NVMe over Fibre Channel。

退后一步，考虑一下 SCSI/FCP 如何支持多路径。在企业级存储阵列上，主流存储阵列控制器架构的设计是为了同等优化所有通往单个 LUN 的路径（不论使用的是哪个控制器目标端口），进而提供对称型多路径。问题在于，很多中端存储阵列能够跨越两个存储阵列控制器双活访问单个 LUN，但是事实上在任意一个已知时间点都只有其中一个控制器拥有 LUN。结果就是，通过一个控制器（即，拥有 LUN 的控制器）的路径被视为首选的优化型路径，而通过另一个控制器的路径被视为不可取的非优化型路径。

异步逻辑单元访问（ALUA）的初衷是确保主机使用优化型路径，并且只有当优化型路径不可用或者运转异常无法支持 LUN 访问时才通过非优化型路径发送 IO。ALUA 是一个在 OS 和存储阵列上实施的 SCSI 标准，以确保 OS 始终使用优化型路径，除非优化型路径不可用或者出现故障。

NVMe over Fibre Channel 也有一个类似的标准——异步命名空间访问（ANA），该标准是 NVMe 规范 1.4 的一部分。ANA 协议定义了存储阵列如何将路径和子系统错误传回主机，这样主机就能管理路径，从一条路径切换到另一条路径，进行故障转移。

切记

在您着手实施 NVMe over Fibre Channel 时，如果您已经有一个使用 ANA 的存储阵列，请确认 ANA 在服务器端的 OS 版本上可用。

了解哪些应用将受益

您可能想将一大堆应用迁移到 NVMe over Fibre Channel。但是，哪些应用无法从性能的提升中受益？您在处理需要不断提高 IOPS、降低延迟的企业应用时，经常需要思考这个问题。您最好是还能在参与流程的服务器上解放一些 CPU 周期出来！

这类应用通常是事务性数据库系统，比如 Oracle 和 MS SQL Server，以及 SAP HANA 和 NoSQL DBMS 系统；此外，您可能也在考虑贵企业的自有应用。该领域的早期采用者倾向于聚焦以下应用：侧重于分析端的应用，比如机器学习、人工智能；或者，他们只是聚焦能够在事务性数据库系统上执行实时分析而不影响应用主用途的能力。

未来，随着企业和应用开发人员从 NVMe over Fibre Channel 的大型队列能力中积累经验，并开发新应用或重构应用，以充分发挥 NVMe over Fibre Channel 和存储级内存存储系统的潜能，我们预计未来全球将涌现很多我们现在只能在梦中想象的应用功能。

推动机器学习向 NVMe over Fibre Channel 迁移的因素

下面，我们将简单介绍下推动机器学习向 NVMe over Fibre Channel 迁移的因素：

- NVMe over Fibre Channel 能够分解计算和存储能力以实现独立扩展，让存储能力突破每个服务器的原生能力。
- 市场上将出现更灵活的集群功能，服务器能够访问同一/共享数据集。
- 机器学习应用需要访问 SAN 内存储阵列中的现有数据，比如事务性系统中的数据，以及物联网 (IoT) 或边缘设备生成的海量非结构化数据。
- 机器学习已经成为了一项业务关键型技术，它需要数据保护和/或高可用性，即使机器学习只是基于主数据集副本运行。

知道并非所有结构都是一样的

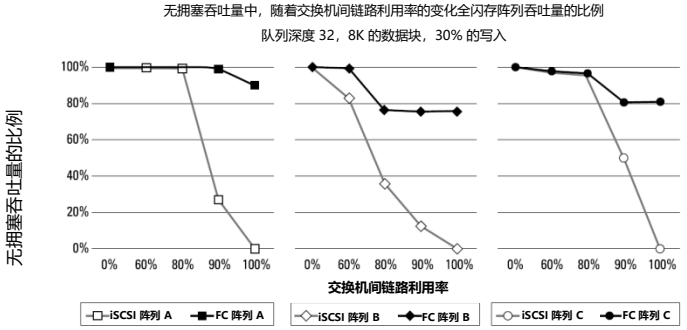
在设计任何网络时，您必须评估在预期的基础架构生命周期内的网络需求，对于数据中心和存储网络来说，通常为未来四到七年的网络需求。在开展评估时，请遵循网络冗余、弹性以及对称/同质拓扑结构方面的最佳实践设计原则，同时规避固有瓶颈。同样，您还需要考虑网络和持续运营的成本，即总体拥有成本 (TCO)。

提示

最佳实践将规划 80% 的网络利用率，同时留出一些空间，以便在不出现性能降级的前提下应对高峰期流量。现实是，网络在荷载情况下如何运转取决于网络设计和使用的网络技术。

Enterprise Storage Group (ESG) 发布了一份报告。在报告中，他们比较了使用 SCSI over Fibre Channel 网络与使用 iSCSI over Ethernet 网络的全闪存阵列 (AFA) 的企业工作负载性能。其中一项测试通过将网络利用率从 60% 提高到 80%、90% 乃至 100%，比较了网络利用率对性能的影响。测试结果展示了光纤通道和以太网（两者的主机和存储一模一样）上网络拥塞的影响。

如图 5-2 所示，当网络上除了受监控的企业工作负载之外没有其他流量时，每个吞吐量性能都实现了归一化。结果显示，当网络出现拥塞时企业工作负载受到了影响，但是光纤通道网络所产生的影响截然不同。当光纤通道网络拥塞程度达到 80%-100% 时，企业工作负载性能会降级近 20%。相比之下，当以太网上的 iSCSI 出现 60% 的拥塞时，企业工作负载性能就会开始降级，然后随着拥塞程度的继续上升，性能会进一步降级，直到拥塞达到 100%，性能完全崩溃。



来源：Enterprise Strategy Group

图 5-2：比较企业工作负载性能。

ESG 报告中的测试结果显示，拥塞对性能的影响纯粹是一个带宽问题。任何网络都有拥塞期，除非不计成本地扩大网络规模。当网络规模适当，能够妥善处理网络上的工作负载时，拥塞期应该只是暂时的，除非因为链路故障或其他组件故障导致性能降级。

针对传输存储流量的网络，您应该知道网络拥塞会带来什么行为或影响，这一点至关重要。当存储流量性能降级时，应用将无法正常运转；当流量中断时，应用将出现更糟糕的表现。结果可能是应用崩溃。高密度事务性系统要从数据库崩溃中恢复过来会相当耗时。与此同时，应用将宕机，企业将停滞不前。

提示

在整个 IT 基础架构生命周期里，您会面临多次网络性能降级。这些问题可能是电缆、光纤和交换机故障或者人为失误所致。在这些事件发生时，高性能的业务关键型应用必须始终可用，并且按照预期运转。

在网络拥塞期维持性能

为什么网络通道饱和时 iSCSI 吞吐量会迅速下降并暂停，而光纤通道却仍能提供大吞吐量？因为这两个结构采用不同的流控制和拥塞控制机制。在光纤通道中，网络负责保证交付，不允许丢包（*无耗*网络），但是 iSCSI 依赖 TCP/IP 来确保实现交付，因为以太网只提供传输，不保证交付。

因此，当网络出现拥塞时，以太网层面会出现丢包，TCP/IP 协议会作出反应，重发丢失的数据包，以尝试适应网络的损耗特性。其中包括接收者向发送者提供的 TCP/IP 数据包确认以及接收窗口（相当于接收者可用的缓冲区空间量）。这一信息将告诉发送者通信两端之间能够传输多少数据。但是，接收窗口仅考虑接收者的缓存区空间，并不考虑任何中介网络节点。因此，随着网络变得拥塞，中介节点可能会耗尽缓冲区空间，开始丢包，这就需要重发数据包。

丢包和重发会引起连锁拥塞，因为重发会占用更多可用带宽，给新的数据块留下更少的吞吐量，甚至有时还会阻碍存储交换的完成。测试结果显示，在最糟糕的情况下，由于丢包和重发占用了所有可用带宽，超时传播到了 SCSI 层，导致 iSCSI 传输终止。

相反，光纤通道则是基于链路到链路、缓冲区到缓冲区的记帐系统运行。连接设备（主机、存储阵列和交换机）后，每个链路两端都能交流可用的缓冲区空间大小。发送者负责跟踪发送者使用了多少链路接收者的缓冲区空间，以及是否有可用的缓冲区用于发送。 每个发送的帧都会使得接收者缓冲区计数相应地递减，每个帧确认都会使得接收者缓冲区计数相应地递减。如果接收者缓冲区计数为 0，则发送者无法发送更多数据。因此，当网络变得拥塞，中介节点可能会耗尽缓冲区空间，引起上游发送者停止发送，进而返回至通信的发起人。

光纤通道的端到端流控制协议包括中介节点，后者将使用公平共享算法，确保每个发送者在缓冲区空间可用后都能公平地获得一定比例的可用吞吐量。因此，即使拥塞程度接近 100%，光纤通道流量也会继续传输，这些测试结果也证明了这一点。


切记

最重要的是，在网络拥塞期间，NVMe over TCP（以太网）的运转方式与 iSCSI 一样（如 ESG 报告所示）。

本章提要

» **您在计划采用 NVMe over Fibre Channel 时必须铭记的十个因素**

第 6 章

有关 NVMe over Fibre Channel 的十大要点

现在，您正式决定采用 NVMe over Fibre Channel。那么，您可以从以下十个要点入手：

» **通过扩展 SCSI/FC 网络以涵盖 NVMe/FC，降低业务运营风险。**安装一个全新的非光纤通道结构基础架构来采用 NVMe，这是一种孤注一掷的方法。此外，您部署哪种基于以太网的协议——iWARP、RoCEv2 还是 TCP？这种方法会给您的长期高价值数据资产以及预算带来风险。您可以选择一种更明智的策略：扩展现有的基础架构，循序渐进地按需实施迁移，以保护您的数据和投资，同时充分利用现有的 IT 技能。

» **借助 NVMe/FC 提高服务器 CPU 利用率和投资回报率 (ROI)。**通过提供精简的 IO 命令、每核多队列和异步提交/完成队列，NVMe 能帮助您大幅提高主机 CPU 利用率。通过提高资源利用率，您可以提高主机和存储资源的利用率，这样，您就可以充分利用您的资源，构建一个可扩展性更高、更经济高效的 IT 基础架构。此外，您还可能发现软件内核许可成本也随之减少。

- **推动更多应用工作负载，因为 NVMe 的 IOPS 将会产生与延迟一样的影响力。**大多数围绕 NVMe 的宣传都是聚焦其无与伦比的低延迟，因为这个衡量指标便于进行基准评估，NVMe 的早期重心放在内存用例上。但是，随着架构向存储和大规模并行处理迁移，IOPS 将变得更加重要，这种情况下，企业愈加需要强大的数据中心结构、分析功能和强大的供应商支持（用于已知的 FC）。

- **立即安装 NVMe 就绪型设备，以便允许应用充分利用下一代低延迟闪存——存储级内存。**这种下一代内存的延迟比当今的闪存技术要低得多。但是，SCM 的成本也更高，为了全面发挥低延迟优势，您只能采用 NVMe 协议。您需要采用混合了 SCM 和传统闪存的分层基础架构，以便优化数据放置，降低基础架构成本。

- **利用多个 NVMe 闪存介质规格，降低 IT 基础架构成本。**NVMe 介质有多种规格，包括插件 PCIe 卡、2.5 英寸的 SFF 驱动器、M.2 和 NF1 扩展组件，以及 NVDIMM 等等。不同的供应商在其解决方案中使用不同规格的介质，优化存储系统性能。请了解技术部署如何影响您的应用，因为这对于您至关重要！

- **让 NVMe 的未来按照其节奏徐徐展开。**与 SCSI 相比，NVMe 协议是一项相对新兴的技术，它于 2014 年问世，未来也将继续快速发展。随着行业逐渐采用闪存技术，业内很快出现了一系列增强包，未来也将有更多增强包随着 NVMe 规范问世。预计，2019 年客户采用率将趋于稳定，NVMe 将在未来几年内取代 SCSI。所有 OS 供应商都将利用 ANA 内置的多路径功能，这是一个重要的发展领域。

- **将供应商支持纳入您的 NVMe-FC 采用计划。**一个可靠的 IT 基础架构有的不只是酷炫的技术；不同供应商的产品也必须能够协同运行。企业供应商对基于 SCSI 的存储的支持非常重要，因为他们将执行全面的可互操作性测试，并且在出现问题时，他们将提供行业领先的支持。IBM 提供覆盖整个数据块、文件和对象存储产品组合的综合端到端 NVMe 存储战略，在这方面，IBM 占据行业领先地位。您必须要求企业测试和支持您的 NVMe 解决方案。因此，请确保询问所有供应商以下事项：他们的 NVMe 存储和结构可互操作性测试、部署计划和后端支持功能。

» **以不同的方式优化光纤通道和以太网/IP。**在以太网和 IP 迅速发展的时代，光纤通道问世并得到了发展。光纤通道之所以能成功，是因为它专为一个主要用例量身定制并进行了优化：可靠、高速地交付突发性*存储*流量。相反，以太网和 IP 作为成功的商品化技术，它们能在任意位置处理数据流量，但是与传统的 SCSI 协议一样，多年以后它们也带来了额外的负担。

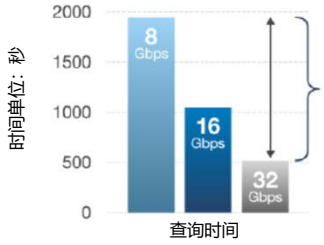
» **单独使用专用存储结构，取得最佳结果。**存储专家知道，任务关键型存储供应商的参考架构需要专用存储结构。正因为此，在处理存储流量时，相比在共享以太网/IP 结构上运行存储，搭载 HBA 的专用光纤通道结构是一个低风险选项，因为前者需要您在支持 RDMA 的 NIC 或 TCP 卸载引擎技术之间进行赌博。为了获得最佳性能和最高的可用性，我们建议您使用面向 FC 和/或以太网的专用存储结构。

» **通过同时运行 SCSI/FC (FCP) 和 FC-NVMe，充分利用光纤通道结构，同时将迁移工作最小化。**通过支持双协议 (FCP/FC-NVMe)，光纤通道结构提供了巨大的优势，因为它支持超低延迟的数据访问以满足工作内存需求；同时它还支持您用低风险方式按需将高价值存储资产从 SCSI 迁移到 NVMe。从 SCSI 迁移到 NVMe 可能需要花费数年的时间，而此类结构还能帮助您简化存储采购决策。

升级至 IBM b-type 32 Gbps Gen 6

利用 NVMe 就绪型存储网络，最大化闪存性能!

升级至 IBM b-type 32 Gbps Gen 6 所带来的优势



71% 4 倍

*www.demartek.com/Demartek_Emulex_LPe32000_Gen6_FC_Evaluation_2016-03.html

如需进一步了解 IBM FlashSystem 和 b-type SAN 解决方案，请访问：

ibm.biz/flashstorage ibm.biz/san-btype

IBM BROCADE A Broadcom Company

展望 *NVMe over Fibre Channel* 前景

NVMe 是当今一种全新且大规模并行的超低延迟内存协议。NVMe over Fibre Channel 将这个突破性的新协议扩展到了企业存储规模。作为一种优质的数据中心结构，光纤通道能够并发传输 NVMe 和 SCSI，为您提供一条低风险的 NVMe 采用路线，保护您的高价值数据资产。本书有助于您快速上手 NVMe over Fibre Channel，帮助您制定采用战略，向您展示前进的道路。

亮点……

- 按照您的步调以低风险的方式采用 NVMe
- 交付速度和可靠性
- 利用增强型队列，提高 IOPS
- 保护高价值资产
- 利用并发 FCP 和 NVMe
- 简化存储采购决策
- 利用 Fabric Vision 开展分析和优化

如需获取相关视频、分步骤演示照片、指南文章或购买本书，**敬请访问 Dummies.com®！**

IBM

BROCADE
A Broadcom Company

Brian Sherman 是一名 IBM 杰出工程师。Marcus Thordal 是 Brocade 的首席解决方案架构师。Kip Hanson 曾经是一名 IT 极客，现在转型为自由撰稿人。他花了大量时间来用浅显易懂的语言阐释难懂的主题，同时对于软盘和 DOS 提示符的没落也经常哀叹不已。

ISBN：978-1-119-60267-5

不得转售

傻瓜系列

还可提供电子版本

Wiley

WILEY 最终用户许可协议

访问 www.wiley.com/go/eula，浏览 Wiley 电子书最终用户许可协议。