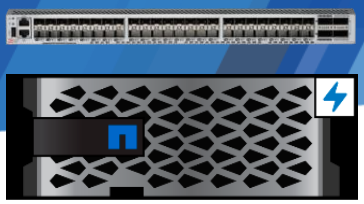
Demartek

NVMe™ over Fibre Channel 的性能优势 — 一种全新的、并行的、高效协议

*相比 SCSI FCP，NVMe™ over Fibre Channel 的 IOPS 提高了**58%**，**延迟低了 34%**。（有什么理由不喜欢？）*



执行摘要

NetApp ONTAP 9.4 是第一个真正意义上可用的企业存储产品，可提供完善的 **NVMe™ over Fibre Channel (NVMe/FC)** 解决方案。NVMe/FC 解决方案基于最近的 T11 INCITS 委员会 **FC-NVMe** 块存储标准，该标准列明了如何按照 NVM Express™ 制定的 NVMe over Fabrics™ (NVMe-oF™) 准则扩展 NVMe over Fibre Channel 命令集。

光纤通道**专为存储设备和系统而建**，是企业数据中心存储区域网络 (SAN) 的事实标准。光纤通道以无损方式运行，配备硬件卸载光纤通道适配器，以及基于硬件的拥塞管理，提供可靠的、基于信用量的流量控制和传递机制，符合 NVMe/FC 的技术要求。

现在的光纤通道适配器还有另一个优势，就是可以运行在相同的适配器、光纤通道网络和企业全闪存阵列 (AFAs) 中**同时使用** SCSI 命令集与 NVMe over Fibre Channel 命令集的传统光纤通道协议 (SCSI FCP)。NetApp AFF A700s 是第一个在同一个端口上同时支持 SCSI FCP 和 NVMe/FC 的阵列。只需要对软件进行简单升级即可实现，保护了现有光纤通道适配器的**投资**，为 NVMe/FC 提供了性能优势。现代光纤通道交换机和主机总线适配器 (HBA) 已可同时支持传统 SCSI FCP 和 NVMe/FC。

在这份测试报告中，Demartek 与 NetApp 和 Broadcom（Brocade 和 Emulex 部门）合作，演示了 NVMe over Fibre Channel 在 NetApp AFF A700s、Emulex Gen 6 光纤通道适配器以及 Brocade Gen 6 光纤通道 SAN 交换机上的优势。

主要发现和结论

- ***NVMe/FC 支持新的 SAN 工作负载：***大数据分析、物联网 (IoT) 和人工智能/深度学习都将从 NVMe/FC 更快的性能和更低的延迟中受益。
- ***NVMe/FC 加速现有工作负载：***Oracle、SAP、Microsoft SQL Server 等企业应用可以立即利用 NVMe/FC 的性能优势。
- ***测试结果：***在我们的测试中，我们发现***在相同的硬件***上，相比 SCSI FCP，NVMe/FC 的 **IOPS 提高了 58%**。我们也观测到了最小的差异，取决于测试，NVMe/FC 的延迟降低了 11% 到 34%。
- ***NVMe/FC 易于采用：***我们观测到的所有性能提升通过软件升级均可实现。
- ***NVMe/FC 保护您的投资：***我们观测到的优势基于支持 32GFC 的现有硬件。
- ***NVMe/FC 数据中心合并：***随着 IOPS 密度的增加，可以在相同的硬件足迹中完成更多工作。

NVMe™ over Fibre Channel 的性能优势 — 一种全新的、并行的、高效协议

Demartek
May 2018

什么是 NVMe over Fibre Channel？

NVMe over Fibre Channel 是一种由两个标准定义的解决方案：NVMe-oF 和 FC-NVMe。NVMe-oF 是 NVM Express 组织制定的一种规范，它与传输协议无关，而 FC-NVMe 是一种 INCITS T11 标准。这两个标准联合定义了 NVMe 如何利用光纤通道。NVMe over Fibre Channel 设计为向后兼容现有光纤通道技术，支持使用相同的硬件适配器、光纤通道交换机和企业 AFAs 的传统 SCSI 协议和新 NVMe 协议。

专为存储而建

光纤通道存储架构可提供一致和高度可靠的性能，是一个独立、专用的存储网络，完全隔离存储流量。光纤通道架构有一个内置的、经**验证的方法，可以发现主机启动器和存储设备**及其在架构上的属性。这些设备可以是启动器，例如带有 FC 主机总线适配器 (FC HBA) 的主机应用服务器和存储系统，也称为存储目标。

对当今的企业数据中心来说，快速访问数据至关重要。传统光纤通道架构的部署通常使用支持**多路径 I/O** 的冗余交换机和端口，以便在出现链接失败时，可以使用备用路径，保持对数据的持续访问。NVMe/FC 也支持多路径 I/O，并通过添加**不对称命名空间访问** (ANA) 来支持首选路径。ANA 添加到了 NVMe 规范中，并于 2018 年 3 月作为技术提案 (TP 4004) 获得批准。这同时要求启动器和目标来实现 ANA。Demartek 认为，今年的一些 NVMe 解决方案将提供首选路径支持（通过 ANA 机制）。

注意：ANA 只适用于 NVMe——其他存储协议有自己的方法来实现多路径和首选路径支持。

光纤通道架构中使用的技术至少可向后兼容前两代产品。这为组织的关键数据资产提供了长期的**投资保护**，并有助于长期的资本预算规划。

光纤通道架构支持多种协议，包括同时支持 NVMe over Fibre Channel 和 SCSI over Fibre Channel。这让组织可以在其当前服务器上使用 Emulex 光纤通道卡、Brocade 光纤通道交换机和 NetApp 全闪存阵列轻松部署 NVMe over Fibre Channel。

为什么要迁移到 NVMe over Fibre Channel？

大多数企业数据中心使用光纤通道 SAN 来存储关键任务数据。运行这些数据中心的许多客户已经拥有运行 NVMe/FC 所需的硬件，包括光纤通道交换机、适配器和存储设备。对于本测试，使用现有硬件迁移到 NVMe/FC 只需要对主机启动器和存储目标进行软件升级。因为 SCSI FCP 和 NVMe/FC 可以同时在同一条线上运行，因此可以根据需要创建 NVMe 命名空间来替换现有的应用 SCSI LUN，并且应用可以引用 NVMe 命名空间来立即获得性能优势。

NVMe/FC 的优势 — NetApp 存储系统

在本测试中，性能改进的最大贡献来自于向存储阵列添加 NVMe over Fibre Channel—***更快的 AFAs 是主要的性能优势***。因为 NVMe 比旧协议更高效，因此 NVMe/FC 架构的优势有很多。这些优势与架构上传输的流量有关，而与通过 NVMe/FC 连接的存储系统内的存储设备类型无关。

NetApp 的 ONTAP 9.4 包括一些新特性，涉及冷数据的自动云分层、对 30TB SSD 的支持以及新的遵从性和安全特性，包括 GDPR 遵从性。但本报告强调的主要新特性是对 NVMe/FC 的支持。

NVMe™ over Fibre Channel 的性能优势 — 一种全新的、并行的、高效协议

Demartek
May 2018

IOPS 优势

更高效的命令集可以交付更高的 IOPS。在我们的测试中，通过简单地从传统 SCSI FCP 命令集中转移到 NVMe/FC，我们发现 IOPS 提高了58%。

延迟优势

与传统 SCSI FCP 相比，NVMe/FC 具有更低的延迟。我们也观测到了最小的差异，取决于测试，NVMe/FC 的延迟降低了 11% 到 34%。

支持现有硬件，性能更高

NetApp 通过简单地向 A700s 应用软件升级许可即可实现这些优势。通过使用相同的存储硬件迁移到 NVMe/FC，可以获得显著的性能提升。后端固态硬盘使用现有接口。

NVMe/FC 的优势 — FC 交换机

Brocade Gen 6 光纤通道架构同时传输 NVMe 和 SCSI (SCSI FCP) 流量，具有相同的高带宽和低延迟。总的来说，NVMe 的性能优势在于终端节点 — 启动器和目标。NVMe/FC 提供的安全性与传统光纤通道协议多年来提供的安全性相同。光纤通道为 NVMe/FC 和 SCSI FCP 提供全架构服务，如发现和分区。最后，NVMe over FC 是第一个 NVMe-oF 传输，符合与 SCSI over FC 相同的高要求，并以全矩阵测试作为使能器，对于企业级支持必不可少。

Brocade 交换机包括 ***IO Insight***，它通过集成的网络传感器主动监控 I/O 的性能和行为，以深入洞察问题并确保服务水平。这一功能是以非干扰和非入侵性地方式从 Gen 6 光纤通道平台上的任何设备端口收集 SCSI 和 NVMe 流量 I/O 统计数据，然后在一个直观的、基于策略的监控和警报套件中应用这些信息来配置阈值和警报。

NVMe/FC 的优势 — FC HBA

本报告中的测试数据体现了 NVMe over Fibre Channel 针对完整解决方案的性能提升。为了更好地解释 NVMe over Fibre Channel 的性能优势，它有助于描述服务器上工作负载的性能改进。NVMe over Fibre Channel 为块存储带来了本地并行性和高效率，而 SCSI FCP 无法做到这一点，并为应用工作负载提供了有意义的性能改进。我们回顾了 Broadcom（Emulex 部门）的测试结果。

当测试启动器性能的特性时（如最大 IOPs），必须使用一个非常快的目标或多个目标，以消除任何可能扭曲测试结果的瓶颈。

数据显示的结果如下：

- NVMe 的目标端效率使单个启动器可以实现超过 100 万 IOPS，且目标比 SCSI FCP 目标更少。
- 工作负载适中的 4KB I/Os 上的 IOPS 改进达 2 倍。
    - PostgreSQL 事务率改进达 2 倍
    - 延迟减少 50% 或更多
    - 当标准化到 CPU 利用率时，IOPS 至少增加 2 倍

NVMe™ over Fibre Channel 的性能优势 — 一种全新的、并行的、高效协议

测试配置 — 硬件

本部门描述本次研究的服务器、存储和存储网络配置。值得注意的是，尽管配置中的所有元素都能够**同时**支持 NVMe/FC 和 SCSI/FC，但在本研究中，它们进行了分别配置，以简化对一个协议的特定参数的修改和优化，而不会影响另一个协议的行为。

服务器（4 台）

- 富士通 RX300 S8
- 2 个 Intel Xeon E5-2630 v2, 2.6 GHz, 6c/12t
- 256 GB RAM (16x 16GB)
- BIOS V4.6.5.4 R1.3.0 for D2939-B1x
- SLES12SP3 4.4.126-7.ge7986b5-默认

光纤通道交换机

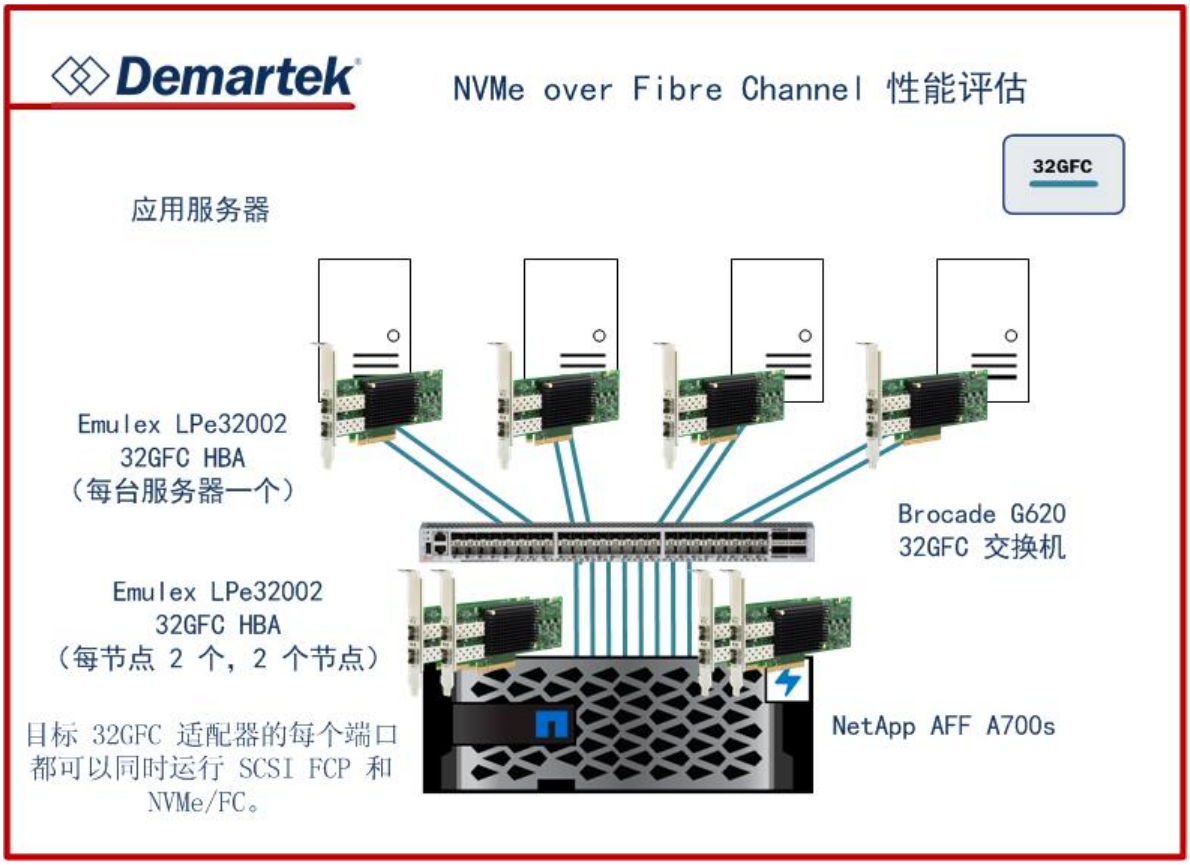
- Brocade G620，48 端口，32GFC
- FOS 8.1.0a

存储系统

- NetApp AFF A700s
- ONTAP 9.4
- 两个节点上各 4 个目标端口，32GFC
- 24 个 SAS SSD，每个 960 GB

光纤通道 HBA

- Emulex LPe32002 32GFC，支持 SCSI FCP 和 NVMe/FC
- 固件版本：11.4.204.25
- 驱动程序版本 11.4.354.0



测试方法

我们的测试目的是比较 NVMe/FC 和 SCSI FCP 在 AFF A700s 存储系统上的性能指标。评估存储系统的最大总体 IOPS 不是本研究的重点。下面几个部分介绍了运行一组合成工作负载时用于衡量这两个协议性能的测试方法和设计考虑。

在我们的研究中，我们通过一个 Brocade G620 网络交换机，为一个 A700s 两节点 HA 存储控制器配置了 4 个运行 SUSE Enterprise Linux 12.3 的服务器。

我们测试床中的 A700s 存储控制器包含两个存储节点。出于测试目的，一个存储节点用于托管 NVMe/FC 容器的存储，一个存储节点用于 SCSI FCP 容器。本测试的设计旨在保证每个协议的全部性能。

表 1 提供了 NetApp 存储控制器配置的详细信息。

存储系统 Active Pair	AFF A700s 配置为高度可用 (HA) active-active-pair
ONTAP 版本	ONTAP 9.4（预发布）
每个节点的驱动器总数	24
硬盘大小	960GB
驱动方式	SAS SSD
SCSI FCP 目标端口	4 个 32GB 端口
NVMe/FC 目标端口	4 个 32GB 端口
以太网端口	4 个 10GB 端口（每个节点 2 个）
以太网逻辑接口 (LIF)	4 个 1GB 管理 LIF（每个节点 2 个，连接到独立的 VLAN）
FCP LIF	8 个 32GB 数据 LIF

测试期间，在给定的时间内，只有一个协议和工作负载是活跃的。注意，尽管参与本测试的每个组件（服务器、HBA、交换机和 AFF A700s）都能够同时支持 FC-NVMe 和 FC-SCSI 的生产流量，但测试中两者被隔离，以便为每个协议收集独立指标，并简化对每个协议的独立特定参数的调整。

我们在 ONTAP 中分别在两个存储节点上创建了一个聚合，分别命名为 NVMe_aggr 和 FCP_aggr。每个聚合消耗 23 个数据分区，横跨 24 个 SAS 连接的 SSD 中的 23 个，为每个数据聚合留出一个分区备用。

NVMe_aggr 包含 4 个 512GB 命名空间。每个 512GB 的命名空间都映射到一个 SUSE 主机来驱动 IO。每个命名空间都包含在自己的 FlexVol 中。每个命名空间都与自己的子系统相关联。

FCP_aggr 包含 16 个 LUN，每个 LUN 都包含在自己的 FlexVol 中。容器总大小与 NVMe 命名空间相同。每个 LUN 都被映射到四个 SUSE 主机上，以便均匀地接收 IO 流量。

我们使用 Vdbench 负载生成工具，针对 A700s 存储目标生成混合工作负载。Vdbench 是 Oracle 提供的一种开源工作负载生成器，可以在 http://www.oracle.com/technetwork/server-storage/vdbench-downloads-1901681.html 中找到。Vdbench 生成各种混合 IO，包括小的随机 IO、大的顺序 IO 和设计用于模拟真实应用流量的混合工作负载。

我们首先执行了一个初始写入阶段来填充精简配置的 LUN 和命名空间。这个阶段通过各个 LUN/命名空间只写入一次，且数据不为零。这确保我们不会读取 LUN 或命名空间中未初始化的部分，这些部分可以在未经适当处理的情况下从 A700s 中得到满足。

我们设计了 Vdbench 工作负载来突出显示一系列用例。这些用例对性能进行了总体概括，并演示了 ONTAP 9.4 中的 SCSI FCP 和 NVMe/FC 之间的性能差异。

1. 合成“四角”测试：16 个 Java 虚拟机 (JVM)， 128 个 SCSI FCP 线程，512 个用于 NVMe/FC 的线程
    a. **大型顺序读取 (64K)**
    b. 大型顺序写入 (64K)
    c. **中型顺序读取 (32K)**
    d. 中型顺序写入 (32K)
    e. **小型随机读取 (4K)**
    f. 小型随机写入 (4K)
    g. 混合随机读取和写入 (4K)
2. 模拟 Oracle OLTP 工作负载：16 个 JVM，100 个线程
    a. **80/20 8K 读取/写入混合**
    b. 90/10 8K 读取/写入混合
    c. **80/20 8K 读取/写入混合，和一个模拟重做日志的独立 64K 顺序写入流**

注意：性能结果提供在上面的**粗体文本**项目中。

工作负载设计

我们使用 Vdbench 5.04.06 和 Java 1.8.0_66-b17 针对 SCSI FCP 和 NVMe/FC 存储驱动不同的混合 IOPS 。这些混合包括使用运行 80/20 select/update 混合的 Oracle 12c 数据库的存储负载配置文件，对 SLOB2 工作负载进行模拟。我们还包括了其他合成 IO 模式，以提供 SCSI FCP 和 NVMe/FC 之间性能差异的一般指示。

注意：我们在这些测试步骤中小心地模拟了真实的数据库和客户工作负载，但是我们承认不同数据库的工作负载是不同的。此外，这些测试结果获取自封闭的实验室环境，没有处于相同基础设施上的竞争性工作负载。在典型的共享存储基础设施中，其他工作负载会共享资源。您的结果可能与本报告中的结果不同。

网络设计

这一部分介绍了用于测试配置的网络连接细节。

网络图显示 FCP SAN 配置了 Brocade G620 32GB FCP 交换机。每个存储节点有四个连接到 FCP 交换机的端口。每个服务器有两个连接到交换机的端口。在测试中，网络连通性没有造成瓶颈。

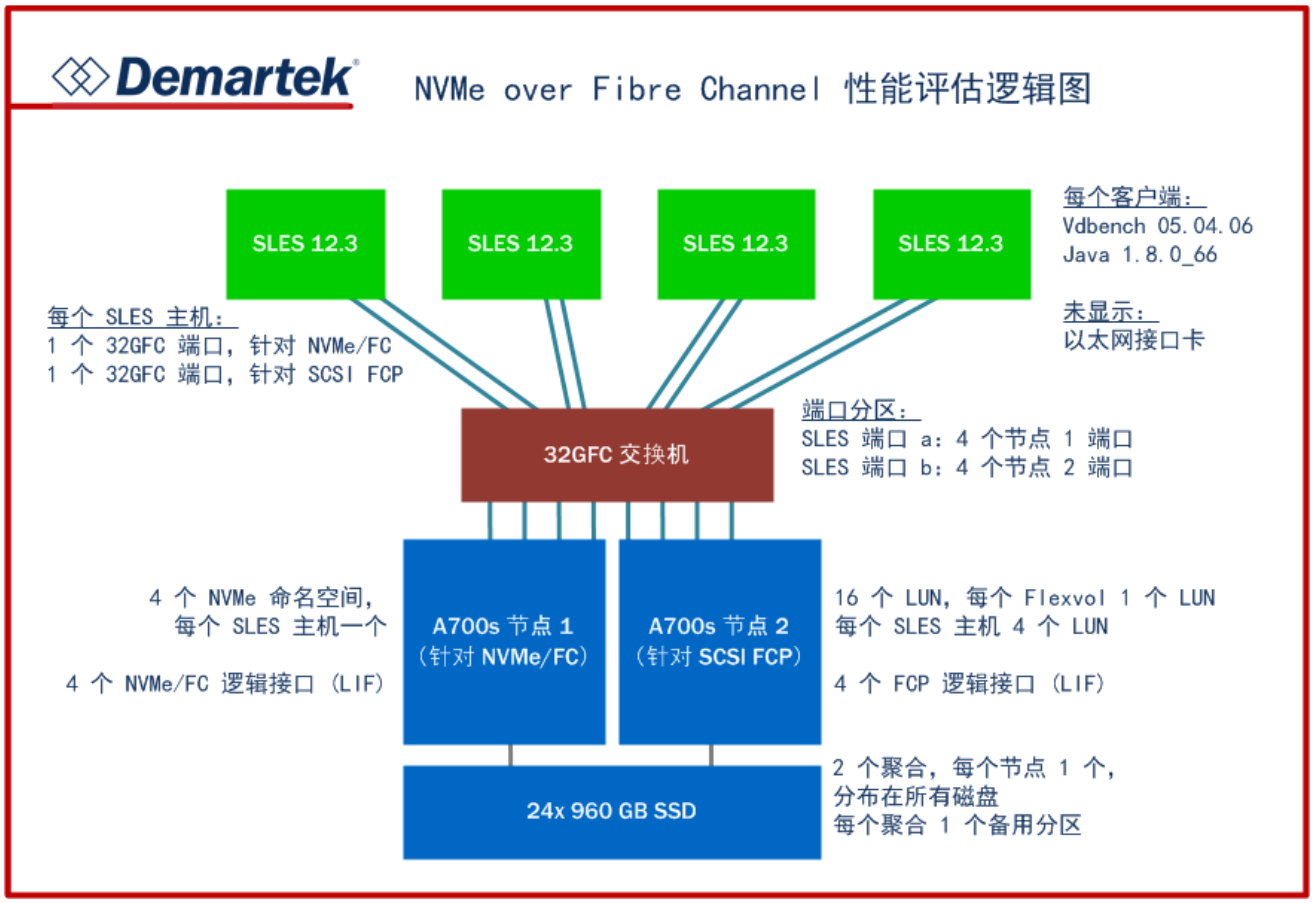
对于以太网连接，每个主机都有一个用于外部访问的 1GBps 链接，并管理节点间的 Vdbench 协调。

我们对每个服务器使用一个 igroup 来包含 FCP 启动器。然后，我们使用“延迟性能”调优配置文件来管理 SUSE 主机。我们手工修改了 FCP DM 设备，以使用“deadline”调度器来提高 SCSI FCP 的性能。

四个 SUSE 服务器各自有一个同时支持两个协议的双端口 FC HBA。两个端口都与 Brocade 交换节连接。每个 A700s 节点有四个连接到同一个交换机上的 FC 端口，总共有八个连接的端口。Brocade 交换机配置有端口分区，将每个 SUSE 主机的端口 1 映射到 A700s 存储节点 1 的所有四个端口。同样，将每个 SUSE 主机的端口 2 映射到 A700s 存储节点 2 的所有四个端口。

NVMe™ over Fibre Channel 的性能优势 — 一种全新的、并行的、高效协议

测试环境逻辑图



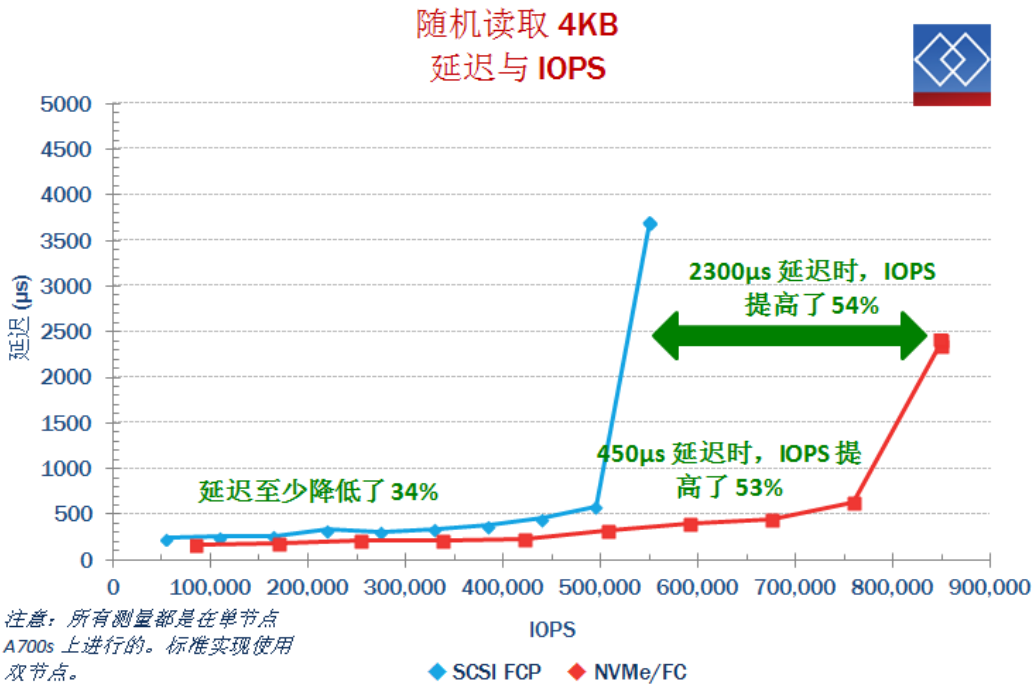
NVMe™ over Fibre Channel 的性能优势 — 一种全新的、并行的、高效协议

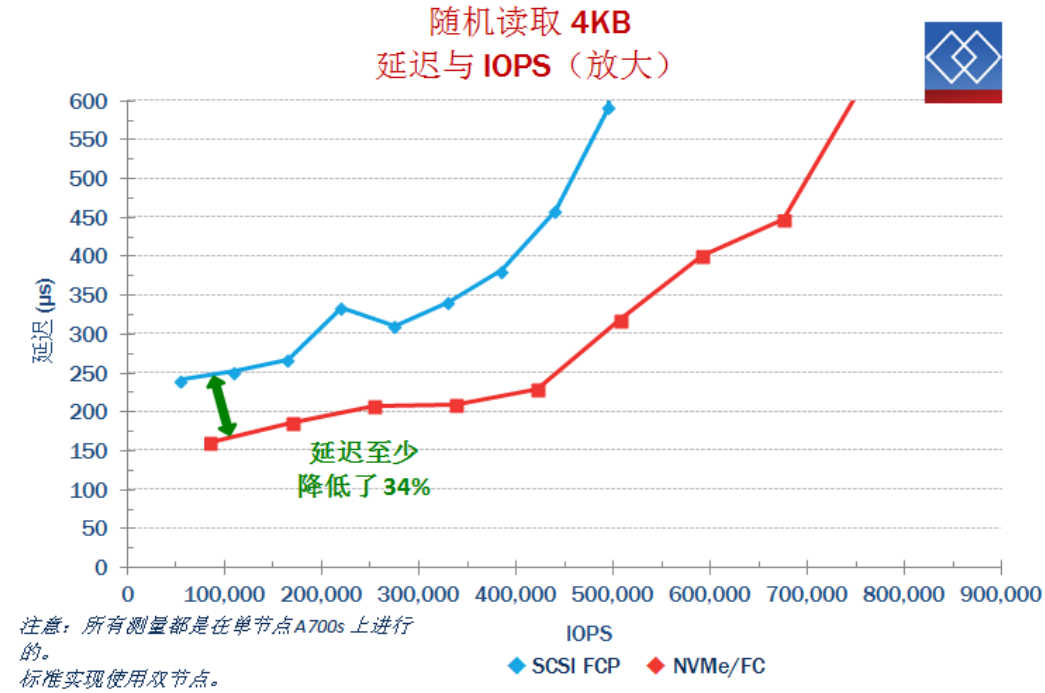
性能结果

所选结果显示在本页和下面两页。所有测量都是在单节点 A700s 上进行的。标准实现使用双节点配置。

随机读取 4KB

对于 4KB 随机读取，NVMe/FC 在 450µs 延迟条件下的 ***IOPS 提高了 53%***。对于 NVMe/FC，延迟至少降低了34%（更好）。本页上第二图表是低于 600µs 的延迟“放大图”。



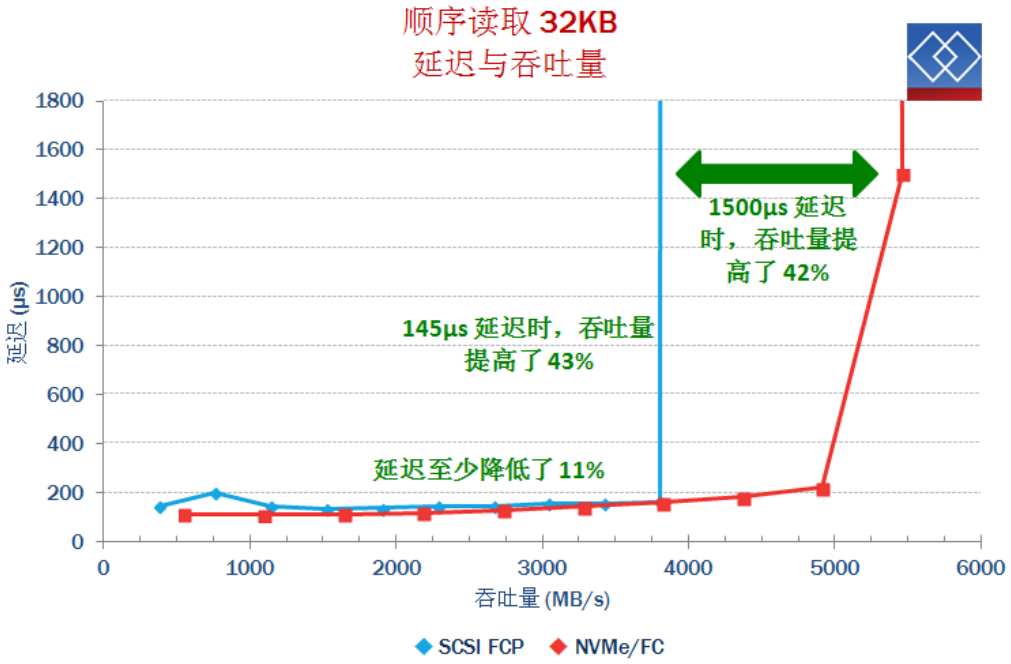


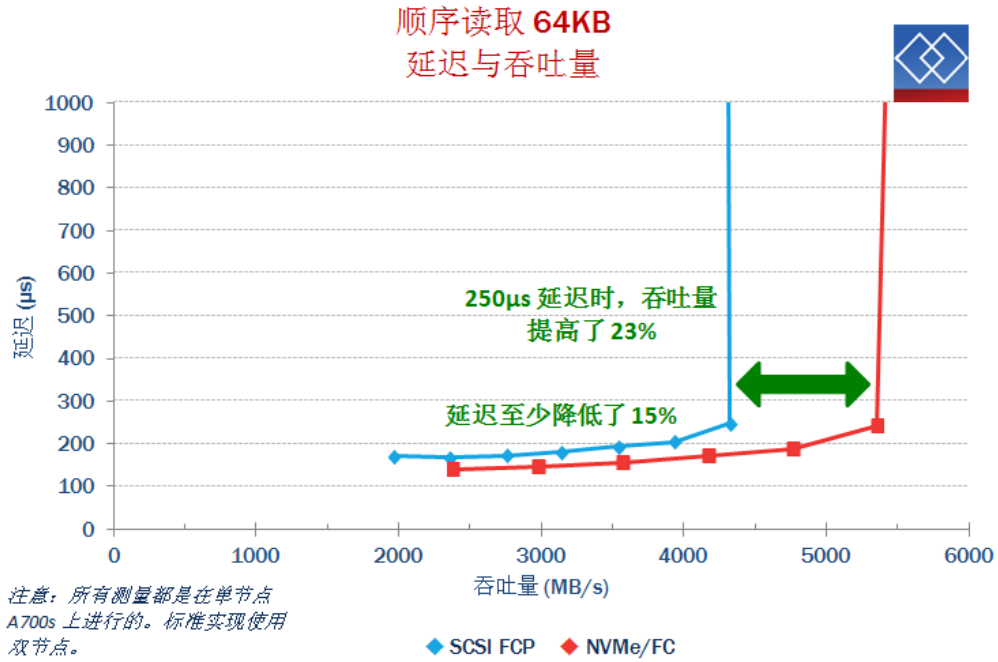
NVMe™ over Fibre Channel 的性能优势 — 一种全新的、并行的、高效协议

顺序读取：32KB 和 64KB

对于 32KB 块大小条件下的顺序读取，NVMe/FC 在 145µs 延迟条件下的 IOPS 提高了 43%。对于 NVMe/FC，延迟至少降低了 11%。

对于 64KB 块大小条件下的顺序读取，NVMe/FC 在 250µs 延迟条件下的 IOPS 提高了 23%。对于 NVMe/FC，延迟至少降低了 15%。





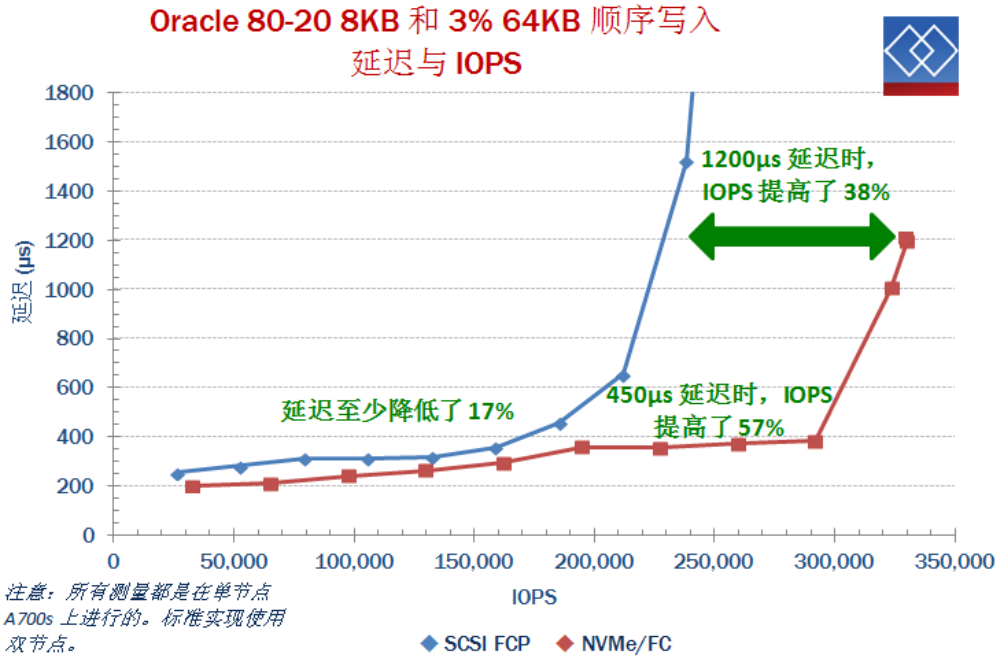
注意：所有测量都是在单节点 A700s 上进行的。标准实现使用双节点。

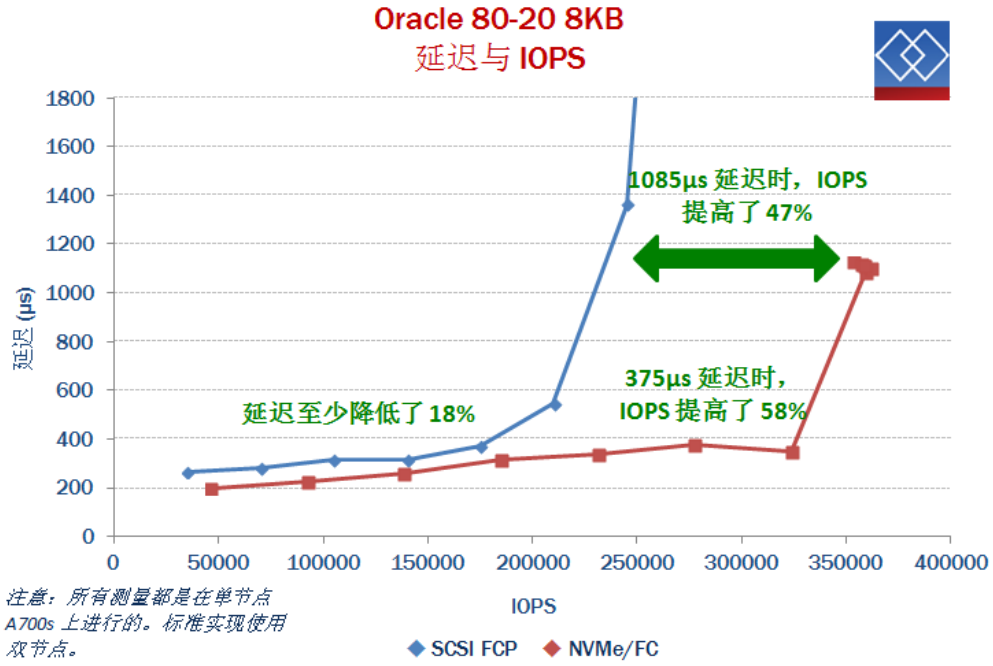
NVMe™ over Fibre Channel 的性能优势 — 一种全新的、并行的、高效协议

模拟的 Oracle 80-20 8KB 工作负载

对于模拟的 Oracle 工作负载，在 8KB（典型的 OLTP 数据库 I/O）和少量 64KB 顺序写入（典型的重做日志）条件下进行 80/20 读取/写入混合，NVMe/FC 在 450µs 延迟条件下的 ***IOPS 提高了 57%***。对于 NVMe/FC，延迟至少降低了 17%。

对于模拟的 Oracle 工作负载，在 8KB（典型的 OLTP 数据库 I/O）条件下进行 80/20 读取/写入混合，NVMe/FC 在 375µs 延迟条件下的 ***IOPS 提高了 58%***。对于 NVMe/FC，延迟至少降低了 18%。

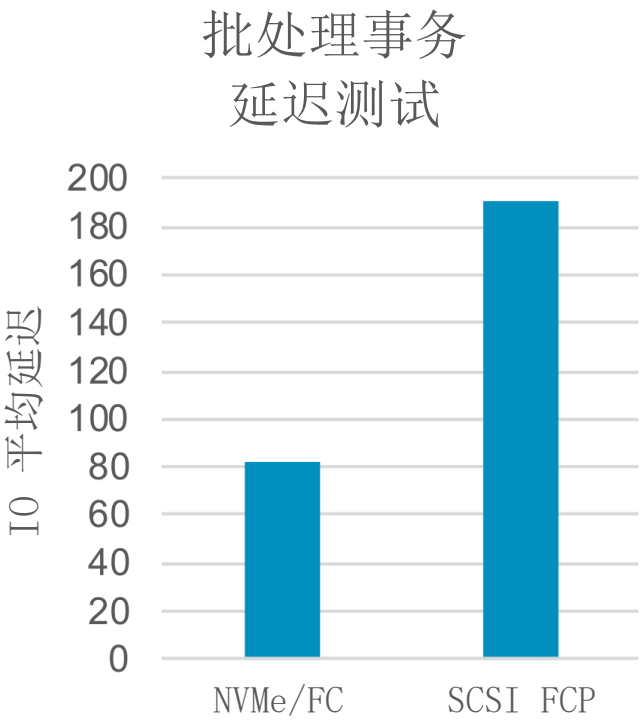




NVMe™ over Fibre Channel 的性能优势 — 一种全新的、并行的、高效协议

NetApp 性能展示

在本报告中，我们检查了单个节点在 NetApp AFF A700s 中的性能改进。NetApp 可以为企业客户演示运行 A300 及 ONTAP 9.4 的 NVMe/FC。NetApp 向 Demartek 展示了以下性能数据（4KB 随机读取 IOs、8 个线程和 1 个队列深度）。本 FIO 测试配置模拟了多种类型的工作负载，该示例是批处理事务。



来源：NetApp

来自 NetApp 演示的数据显示，在 NetApp A300 中，其 NVMe/FC 的延迟降低了一半——以前仅在内部 SATA 和 SAS SSD 看到的延迟级别。NetApp 邀请您与您的 NetApp 代表联系，立即安排您的 NVMe over Fibre Channel 演示。

归纳与总结

NVMe/FC 通过光纤通道健全、可靠的企业级存储区域网络技术，充分利用 NVMe 的并行性和性能优势。

在我们的测试中，通过使用 NVMe/FC，我们发现与使用相同硬件的传统 SCSI FCP 相比，IOPS 提高了 58%。对于测试的配置，只需要对主机启动器和存储目标进行软件升级。这意味着可以很容易地采用在光纤通道技术上的现有投资，而不需要购买新硬件。因此由于提供了整合机会，每平方英尺的性能可能有更多的改进。此外，通过采用 NVMe/FC，可能有机会延迟购买新服务器和存储硬件，从而节省潜在的硬件和软件许可成本。

NVMe/FC 使现有应用的性能得到了提升，并使组织能够利用现有的基础设施处理高要求的新应用，如大数据分析、物联网和人工智能/深度学习。对于测试的配置，只需要对主机启动器和存储目标进行软件升级，这些都可以实现。这使得 NVMe/FC 让组织可以按自己的节奏采用，无需叉车式升级或学习全新架构技术的所有细节。

Demartek 认为，NVMe/FC 是一种（可能显而易见的）优秀的技术，尤其对于已经拥有光纤通道基础设施的组织，如果正在考虑 NVMe over Fabrics，是采用光纤通道技术的好理由。

本报告的最新版本可以从 Demartek 网站上获取：
https://www.demartek.com/Demartek_NetApp_Broadcom_NVMe_over_Fibre_Channel_Evaluation_2018-05.html 。