NVMe/TCP 对比 iSCSI

Version 1.0

August 2018

目录

1. 总览

当前，在许多情况下，iSCSI是用于通过TCP / IP进行块存储的标准协议。

Lightbits Labs™ 和 NVM Express 工作小组 正在就高性能，低延迟块访问的新标准NVMe / TCP 进行协作。

本文介绍：

- iSCSI 和 NVMe / TCP之间的主要架构差异。
- 为什么基于TCP / IP的NVMe是 分离闪存 (disaggregated flash storage)的最佳解决方案
- 性能数据证明了NVMe / TCP是架构改进的重要性

2. iSCSI 背景

自1990年代初以来，小型计算机系统接口（SCSI）协议一直是块存储设备与非消费者计算机 和 服务器接口的主要标准。

最初的SCSI协议基于共享总线传输。 最终，由于增加的带宽和距离要求，开发了串行SCSI传输。 这些包括：

- Fibre Channel (FC) 光纤信道
- Serial Attached SCSI (SAS) 串行连接的SCSI
- SCSI over TCP/IP (iSCSI)

由于TCP / IP的广泛使用，iSCSI已成为标准网络上块存储的通用标准。 但是，所有这些协议不过是为旋转硬盘驱动器开发的相同原始SCSI协议的传输方式。

另一方面，闪存与旋转驱动器的行为大不相同。 闪存不使用序列化数据访问的“单头”。 对闪存的数据访问可以高度并行化，这是NVMe规范背后的主要推动力。 该规范最初旨在通过PCI Express（PCIe）接口存储设备。

在制定了NVMe规范之后，NVMe工作组开发了NVMe over Fabrics协议，以支持通过网络而不只是PCIe的NVMe协议。 在此规范中最初定义的传输是使用RDMA的光纤通道和以太网。 这些传输需要专用但不常用的数据中心设备。 因此，就像在标准以太网和TCP / IP上启用iSCSI的SCSI一样，NVMe / TCP对于NVMe也是如此。

3. The iSCSI Queueing Model 排队模型

在SCSI协议中，客户端是“发起者”，访问的存储实体是“逻辑单元”。在存储系统中，逻辑单元与协议不可知术语“卷”同义。发起者之间的访问模型 逻辑单元由命令队列组成，其中启动器将要由逻辑单元执行的命令排队，并为每个排队的命令接收数据和确认。 该体系结构的核心是每个启动器-逻辑单元连接都存在单个队列。

由于iSCSI只是通过TCP / IP的SCSI传输，因此此概念直接转换为由单个线程打开的单个TCP / IP套接字组成的发起程序逻辑单元连接。 由于SCSI是无处不在的存储协议，因此操作系统在其存储堆栈中实现了相同的模型。 对存储卷的访问始终由单个队列来调解，该队列对应于启动器与逻辑单元的连接。

这种单一队列模型几十年来没有出现任何问题。 但是，由于典型的服务器具有数十个CPU内核，因此当今的系统呈现出不同的现实。 因此，由多个内核并行执行的存储操作需要通过单个存储队列上的锁来进行中介。 这对于旋转仅执行约100 IOPS的硬盘驱动器或每卷执行几千IOPS的典型企业存储系统来说不是问题。

但是，对于闪存存储而言，单个队列模型存在问题，因为单个SSD能够存储数十万个IOPS。 在具有多个内核的服务器上，队列争用成为严重的性能问题。

在iSCSI的情况下，单队列模型进一步受制于单线程和单TCP套接字的标准实现。 通过通过多个目标端口显示逻辑单元并在多个线程和套接字上实现负载均衡I / O，可以执行多路径I / O。 但是，这种策略在目标端管理起来很复杂，并且始终会在发起方端受到排队问题的困扰-因为操作系统仍将多条路径视为单个设备。

4. NVMe 背景

自从成为访问高性能SSD的最先进协议以来，NVMe已经改变了存储行业。

NVMe最初是为高性能直连PCIe SSD设计的，后来通过NVMe over Fabrics（NVMe-oF）进行了扩展，以支持机架规模的SSD远程池。 IT业界已广泛接受，这种新的NVMe-oF模型将取代iSCSI协议作为计算服务器和存储服务器之间的通信标准，并成为用于分解存储和计算服务器的默认协议。

NVMe-oF的初始部署选项仅限于仅适用于小型部署的RDMA（远程直接内存访问）和光纤信道（FC）结构。

4.1. NVMe/TCP 发展历程

NVMe规范已经成为高性能SSD的最新协议。与SCSI，iSCSI，SAS或SATA接口不同，NVMe实现了简化的命令模型，并且针对 多核服务器CPU优化的 (multi-queue)多队列体系 结构。 结构上(NVMe-oF）规范扩展了NVMe以通过 "网络"共享PCIe SSD，最初的实现是使用RDMA结构。

如今，Lightbits Labs正在与Facebook，英特尔(Intel) 和其他行业领导者合作，以扩大NVMe-oF标准支持与RDMA结构互补的 TCP / IP传输。.

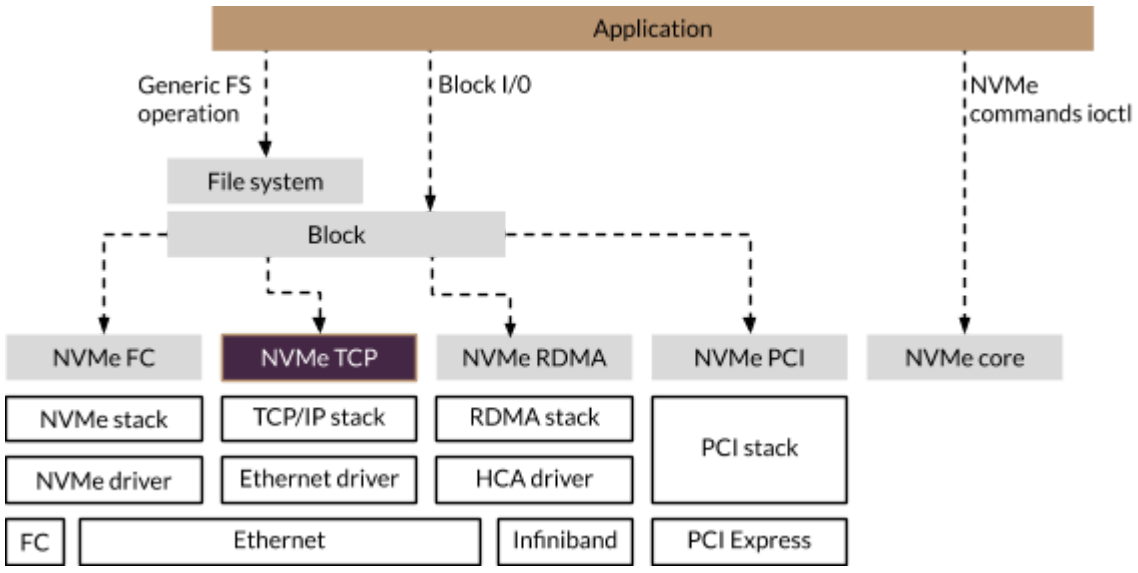
使用NVMe / TCP进行(disaggregated)分离具有明显的优势，简单而高效。 TCP无处不在，高扩展，可靠，并且是短暂连接和基于(container applications)容器应用程序的理想选择.

此外，由Shared Flash共享闪存迁移到 NVMe/TCP存储不需要更改数据中心网络基础架构。无需更改基础架构即可在整个数据中心轻松部署，因为几乎所有数据中心网络都旨在承载TCP/IP。.

业界在NVMe / TCP协议上的广泛合作意味着该协议是从头开始设计的，具有广泛的生态系统支持，并且考虑到了对任何操作系统和NIC（网络接口卡）的支持。 NVMe / TCP Linux驱动程序是Linux Kernel的自然匹配，并且使用标准Linux网络堆栈和NIC，无需进行任何修改。

结果是可望成为超大规模数据中心量身定制的新协议，无需更改基础网络基础架构即可轻松部署.

Figure 1: NVMe/TCP 与Linux内核中的现有NVMe协议无缝集成



4.2. NVMe/TCP 优点

(NVMe / TCP）使用简单高效的TCP / IP结构将NVMe扩展到整个数据中心，并具有以下优点:

- 使跨数据中心(availability zones)可用区 和 (regions)区域的分离
- 利用广泛的TCP / IP传输为高度并行的NVMe软件堆栈 带来较 低的平均延迟和(tail latency)尾延迟
- 无需更改网络基础架构或应用服务器程序
- 提供高性能NVMe-oF解决方案，该解决方案具有与直接连接的SSD（DAS）相同的性能和延迟
- 使用针对NVMe和现有数据中心进行了优化的高效, 简化的 块存储软件堆栈
- 提供对为现今的 多核应用程序/客戶端服务器 优化的存储的并行访问
- 支持基于标准的解决方案，该解决方案将获得行业范围内的支持

5. The NVMe Queueing Model 排队模型and Linux块设备层中的Block Multi-Queue

NVMe规范认识到对固态存储的高度并行化访问，因为不存在单个瓶颈，例如旋转驱动器中的机械头。因此，对NVMe“命称间”的I / O等同于SCSI逻辑单元, 可以通过64,000个队列执行，每个队列最多64,000个命令。.

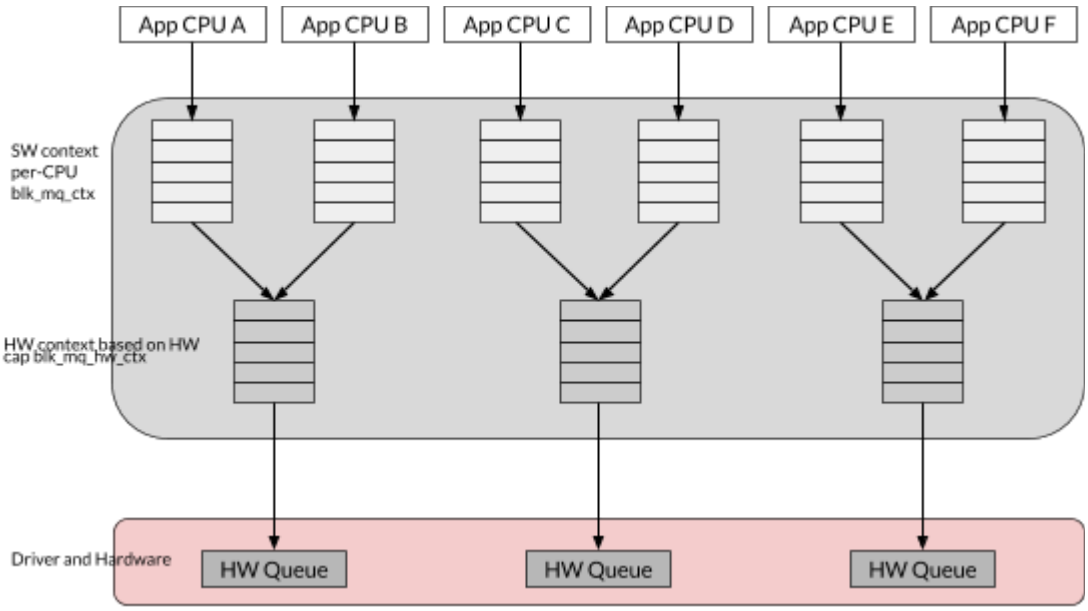
这种并行访问是结构开发的一项关键，因为许多队列消除了SCSI 固有模范的争用。.

5.1. Linux Block Multi-Queue 块设备层

新的Linux Block Multi-Queue堆栈（blk-mq) 开发利用了新的NVMe架构。

blk-mq使运行在 内核上 的(App)应用程序 或 (Thread) 线程 , 接收每个卷映射到软件队列的专用软件队列。**这些队列直接对应于基础硬件中的队列。 因此，如果硬件队列的数量等于核心的数量，则没有队列争用或锁定。**

Figure 2: Linux blk-mq model



5.2. NVMe/TCP and blk-mq

NVMe / TCP协议利用了现代数据中心NIC可以处理不同硬件队列中不同TCP流量的发展。

通过使用blk-mq模型,NVMe / TCP启动器驱动程序 为每个命称空间在每个CPU内核中打开一个TCP套接字，和这些插座映射到不同的硬件队列。从而，与iSCSI不同，这意味着具有高并行度，没有队列争用，并且启用了更高的性能。

6. NVMe/TCP 与 iSCSI 性能对比结果

对于以下基准测试结果, 用于在行业标准FIO基准测试工具iSCSI和NVMe / TCP的相同设置。

- 启动器是双插座 Intel® Xeon® E5-2648L v4 处理器运行在 1.8 GHz, 每个插座具有14个内核，每个内核具有两个(hyperthread)超线程。
- 目标计算机具有相同的插槽和核心数, 及运行于 1.7Ghz 的 Xeon E5-2650L v4处理器
- 目标中的SSD是Intel DC P3520，每个450 GB。
- 所有I / O均通过两个Mellanox ConnectX-4 100GbE NIC执行。

启动器和目标计算机都运行Linux 4.13.0-rc1内核版本。对于iSCSI，使用了内核中的备用iSCSI启动器。对于NVMe / TCP，该测试使用了开源启动器，该启动器将很快由Lightbits公开提供并提供。

对于iSCSI存储，配置了一个12驱动器的RAID-0组，以及Lightbox™NVMe/TCP，exported namespace(导出的名称空间)使用相同的硬件和相同的12个SSD。在运行基准测试之前，所有SSD都被完全覆盖以对测试进行预处理。

对于iSCSI，每个会话的最大命令数设置为2048，和逻辑单元队列深度均为1024-两个参数,是实施允许的最大值。

从而,该测试是对两种具有相同硬件和软件配置的基于TCP的协议的合理比较。该基准测试演示了一个典型的多线程I/O绑定应用程序。这样的应用程序通常会为每个任务打开一个线程，每个线程执行一个I/O(即多个线程，每个线程的I/O队列深度为1）。然后,看的I/O可扩展性线程的数量上升,线程数在变化。改变线程数是应用程序的关键测试,在当今的现代服务器上运行数十个CPU内核。I/O混合是70%的读取和30%的写入。

Figure 3: IOPS scalability for multi-threaded applications
Note: Random 4KB RW70/30 QD=1

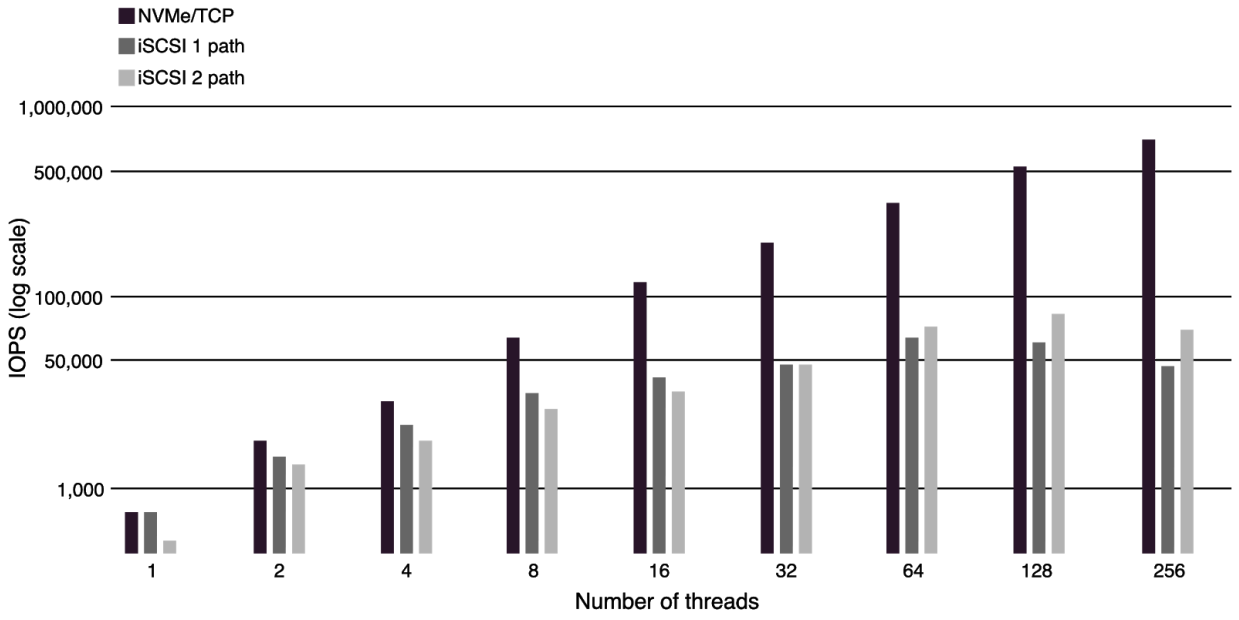
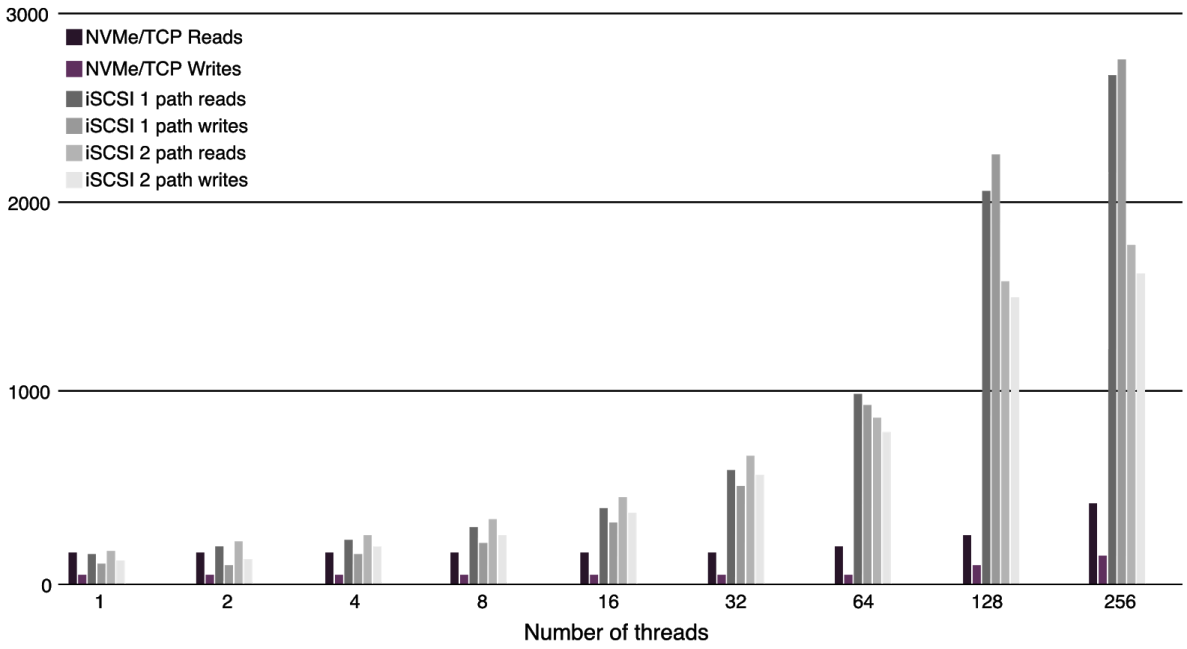


Figure 4: Mean I/O latencies in multi-threaded applications
Note: Random 4KB RW70/30 QD=1



6.1. iSCSI 测试结果

在(active-active)双活目标模式下的两个(multipathed)多路径目标 连接上运行128threads个线程时，iSCSI达到了89K IOPS的峰值。

此IOPS数量是此类应用程序可以获得的最大I / O性能。 同样重要的是要注意，与单个连接相比，在两个目标连接上的多路径具有很小的好处。

同样令人不安的是，超过64个线程的iSCSI会导致平均延迟超过一毫秒。在当今的大规模多核超大规模云服务中，闪存驱动器的I / O性能是无法接受的。

这些发现说明了iSCSI中固有的性能问题：协议的单队列，单线程模型限制了可伸缩性。

6.2. NVMe/TCP 测试结果

NVMe / TCP测试结果显示了本机多线程体系结构的优点。在此特定应用程序场景中，单个卷（NVMe名称空间）的I / O轻松上升到725K IOPS-该应用程序的IOPS高出一个数量级。此外，在256个线程处，启动器或Lightbox™的性能不会达到饱和。 重要的是，NVMe / TCP的I / O延迟远低于具有相似线程数且IOPS较低的iSCSI延迟。 即使在256个线程和725K IOPS时，NVMe / TCP延迟仍低于500s 。

7. 结论

NVMe的开发旨在通过PCIe接口闪存，旨在实现对数据的高效并行访问，特别是对于多核服务器而言。

简单来说，基于SCSI的旧式存储协议无法跟上这一发展。因此，NVMe over Fabric是分散式闪存或闪存SAN的必需开发。

今天，大多数数据中心不支持使用RDMA或光纤通道等技术的NVMe over Fabrics。
这些技术是：

- 仍然离主流太远
- 需要专用设备
- 数据中心网络基础架构的需求发生重大变化
- 将客戶锁定到单个硬件供应商。

因此，该标准的自然发展是通过TCP/IP支持NVMe 。

通常将NVMe / TCP与之前通过iSCSI同级的块存储进行比较。Linux内核已经集成了iSCSI堆栈多年，并且有数十种现成的iSCSI产品。简单的架构分析和基准测试比较表明，使用相同的硬件和可比较的软件时，NVMe / TCP提供的网络性能比iSCSI更高，更快。这些结果表明NVMe / TCP有望成为基于TCP的块存储的领先标准。

Lightbits是NVMe / TCP的发明者和推动力，这将成为主要的网络块存储协议。该协议通过为每个客户端应用程序启用极高的IOPS，释放了分散的闪存存储的潜力，延迟很短。

总体而言，灯箱性能可扩展至超过5百万次IOPS，因此，可以为众多客户提供高性能的高性价比操作。

关于 Lightbits Labs™

当今的存储方法是为企业设计的，无法满足不断发展的云规模基础架构要求。

例如，SAN因缺乏性能和控制能力而广为人知。从规模上讲，直接连接的固态硬盘（DAS）变得过于复杂，无法顺利运行，成本也很高，并且固态硬盘利用率低下。

云规模的基础架构需要对存储和计算进行分解，顶级云巨头从低效的直接连接SSD架构过渡到低延迟共享NVMe闪存架构证明了这一点。

与其他NVMe-oF方法不同，Lightbits NVMe / TCP节省成本的解决方案将存储和计算分开，而不会影响网络基础架构或数据中心客户端。

Lightbits团队成员是NVMe标准的主要贡献者以及NVMe over Fabrics（NVMe-oF）的发起者之一。

现在，Lightbits正在制定新的NVMe / TCP标准。作为该领域的开拓者，Lightbits解决方案已经在行业领先的云数据中心中成功进行了测试。

该公司的共享NVMe架构提供了有效而强大的分解。如此平稳的过渡过程，您的应用团队甚至都不会注意到这一变化。他们现在可以以比本地SSD更好的尾部延迟发狂！

最后，您可以将存储与计算分开，而不会造成麻烦。

www.lightbitslab.com info@lightbitslabs.com

US Office
1830 The Alameda,
San Jose, CA 95126, USA

Israel Offices
17 Atir Yeda Street,
Kfar Saba, Israel 4464313

3 Habankim Street,
Haifa, Israel 3326115