为什么NVMe/TCP 是数据中心的更好选择

非易失性内存表示(自从作为高性能固态驱动器 (的最新协议出现以来，已经改变了存储行业。
最初是为高性能的直接附加PCIe SSD 设计的，后来 NVMe 与 NVMe Over(NVMe o F 进行了
扩展，以支持机架级 r ack scale 的 SSD 远程池。业界普遍认为，这种新的NVMe o F 模式将取代互联网小型计算机系统接口 i SCSI) 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 分离 存储 disaggregated storage 的默认协议。然而，NVMe o F 的初始部署选项仅限于 网状 通道 F ibre Channel 和远程直接内存访问 Remote Direct Memory Access , RDMA) 结构。如果能提供一种新的、更强大的技术来提供 NVMe 的速度和性能，而不需要高昂的部署成本和复杂性 的方案会怎么样 呢？ NVMe over TCP(NVMe/TCP)使用简单有效的 TCP/IP 结构将 NVMe 扩展到整个数据中心。

本文描述了 NVMe/TCP 如何成为现有数据中心的一种更好的技术及其带来的好处。
这些优势包括:

- 支持跨数据中心的可用性区域和区域进行分解
- 利用无所不在的 TCP 传输，具有低延迟、高度并行的 NVME 堆栈
- 无需在应用服务器端 高性能 NVM e oF 解决方案 上进行更改
- 该解决方案提供类似 DAS SSD 的性能和延迟来
- 一种为 NVM e 优化的高效、流线型块存储网络软件栈
- 为当今的多核应用程序 客戶端服务器操作提供了对存储的并行访问
- 标准 NVM e 标准的控制路径

1. NVMe/TCP 概述

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， NVM e实现了为多核服务器 CPU 优化的简化命令模型和多队列体系结构。规范的 NVM e 扩展了NVM e ，在网络上共享 PCIe SSD ，初始实现使用 RDMA 结构。

今天， Lightbits 实验室正在与Facebook 、英特尔和其他行业领袖合作，扩展 NVM e 标准，以支持与 RDMA 织物互补TCP/IP 传输。

基于NVME/TCP 的 分离 集 群 具有简单高效的显著优点。TCP 具有普及性 、可扩展 性 、可靠性， 对于短暂在线连接和基于容器的应用 是一种 理想 的 选择。

此外，迁移到与 NVM e /TCP 共享的闪存不需要更改数据中心网络基础设施。没有基础设施的改变意味着在数据中心之间的部署很容易，因为几乎所有的数据中心网络都是为了携带TCP/IP 而设计的。

在NVM e /TCP 协议上的广泛行业协作意味着该协议是从基础上设计的，具有广泛的生态系统，对任何操作系统和网络接口卡对任何操作系统和网络接口卡都都支持。

NVME/TCP 驱动程序 已成为 Linux 内核的 标准 匹配，使用标准的 Linux 网络堆栈和 NIC ，无需任何修改。 这种 很有前途的新协议为 超规模 数据中心量身定做，在不改变底层网络基础设施的情况下很容易 部署。

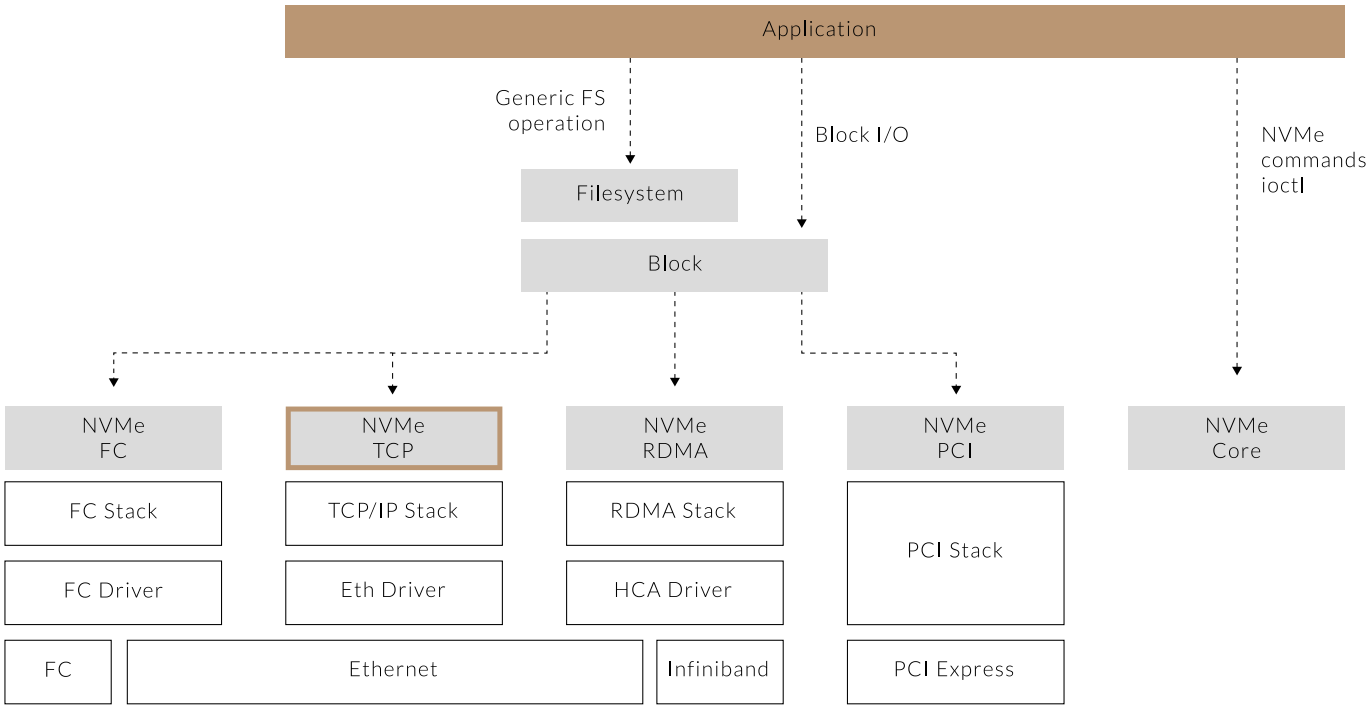

Figure 1: NVMe/TCP seamlessly integrated with existing NVMe protocols in the Linux kernel

2. 今天的数据中心是如何处理存储的

2.1. 直接连接存储结构与 NVMe

NVMe存储协议旨在从固态驱动器 (中提取全部性能。 NVMe 协议中设计的并行性有助于实现这一性能。 NVMe 消除了单队列 iSCSI 模型。 相反， NVMe 在 CPU 子系统和存储之间支持多达 64 000 个队列。

SSD是使用多个并行通信信道与多个 SSD 存储位置连接的并行设备，这意味着 SSD 可以在大型并行流中有效地接收数据。在 NVMe/TCP 协议存在之前，利用这种并行性的最简单方法是将 NVMe SSD 直接安装到应用服务器。换句话说，您必须使用 DAS 模型构建存储基础结构。

使用DAS 方法，应用程序受益于：

- 多个 CPU
- 多个 NVMe I/O 队列
- 并行 SSD 体系结构

业界面临的挑战是将SSD 从单个服务器转移到共享存储解决方案，同时提高基础设施的利用率而不损失 DAS 的性能增益。因此，所有 NVMe 分解技术的目标是在共享 NVMe 解决方案中实现 DAS 性能。

2.2. 前一代基于 IP 的存储体系结构

以前，iSCSI 标准是通过 TCP/IP 网络连接到块存储的唯一选项。它是在世纪之交开发的，当时大多数处理器都是单核设备。在 SCSI 中，应用 启动器 initiator 和存储 目标 target 之间只有一个连接 通过 单一 的 TCP 套接字 socket 将 客戶端 连接到块存储服务器。.

今天，数据中心处理器是大规模并行多线程设备。 当今处理器的这种复杂性要求对可用的存储协议进行彻底改革。其结果是 NVMe 作为 SATA 和 SAS(串行附加 SCSI) 的替代品出现了。

在所有这些早期协议中，它们的开发都是基于一个序列化的、旋转的磁盘驱动器。

非易失性存储器(是一种并行存储技术，它不需要盘片或多个盘片在磁头或磁头下旋转。使用 NVM 存储设备，许多内存位置可以并行访问和延迟较低的位置。

毫无疑问，iSCSI 仍然适用于具有低到中等存储性能需求的用例。然而， iSCSI 不能满足大规模低延迟的 I/O 密集型应用的需求。

NOTE: To learn about head-to-head performance measurements between iSCSI and NVMe/TCP, send an email to info@lightbitslabs.com requesting a copy of Lightbits' whitepaper comparing iSCSI vs. NVMe/TCP.

2.3. 分离NVM e /TCP 的替代办法

RDMA和基于汇聚以太网 (的远程直接内存访问 (以及光纤信道上的NVMe (NVMe Over FC) 也是其他试图解决分解问题的网络存储协议。然而，这些选择要求在两端 应用服务器和存储服务器 安装昂贵的特殊硬件，例如 RDMA 功能的 NIC 。此外，在安装了 RDMA 硬件之后，在 可 RDMA 交换结构中配置和管理流控制也是非常复杂的。RDMA确实提供了适用于某些高性能计算环境的性能，但它需要增加成本，并且需要非常复杂的部署。TCP/IP已被证明在 超大规模 环境中可靠有效地工作。 NVMe/TCP 继承了这种可靠性和效率，可以与 RDMA 作为互补解决方案共存，也可以完全取代 RDMA

3. 数据中心中的分离闪存 和 NV Me/TCP 解决方案

在DAS 环境中， 驱动器 是 部署到服务器 中 或与服务器一起部署之前购买 的 ，并且它们的容量利用率随着时间的推移缓慢长。另外 为了避免耗尽存储所带来的后勤问题， DAS 配置常常被故意过度配置。

相反，将存储从计算服务器中分离出来的数据中心更有效。存储容量可以独立地进行缩放，并且可以根据需要分配给计算服器。随着每GB 闪存成本的降低，分类存储方法更加经济，而且数据中心部署的前期成本要低得多。通过动态分配存储资源，避免了超 配（ over provisioning 开销，大大降低了总体成 本。NVMe/TCP解决方案开启了高性能固态驱动器 (云基础设施的潜力。

它使数据中心能够从低效的直接附加的 SSD 模型转移到一个共享模型，在该模型中，计算和存储是独立的，以最大限度地提高资源利用率和操作灵活性。这种新的共享模式采用了创新的 NVMe/TCP标准。 L ightbits Labs 发明了这一概念，并正在引领这一新标准的发展。

NVMe/TCP不会影响应用的性能。实际上，它经常改善应用尾部延迟，从而改善用戶体验，并使云提供商能够在相同的基础设施上支持更多用戶。它不需要对数据中心网络基础设施或应用软件进 行任何更改。它降低了数据中心的总拥有成本 (TCO)，并使维护和缩放超大型数据中心变得更容易。 Lightbit s 实验室正与其他市场领先者合作，争取在业界广泛采用这一标准。

NVMe/TCP利用标准的以太网拓扑结构，独立地进行计算和存储，以达到最大的资源利用率，并降低 TCO

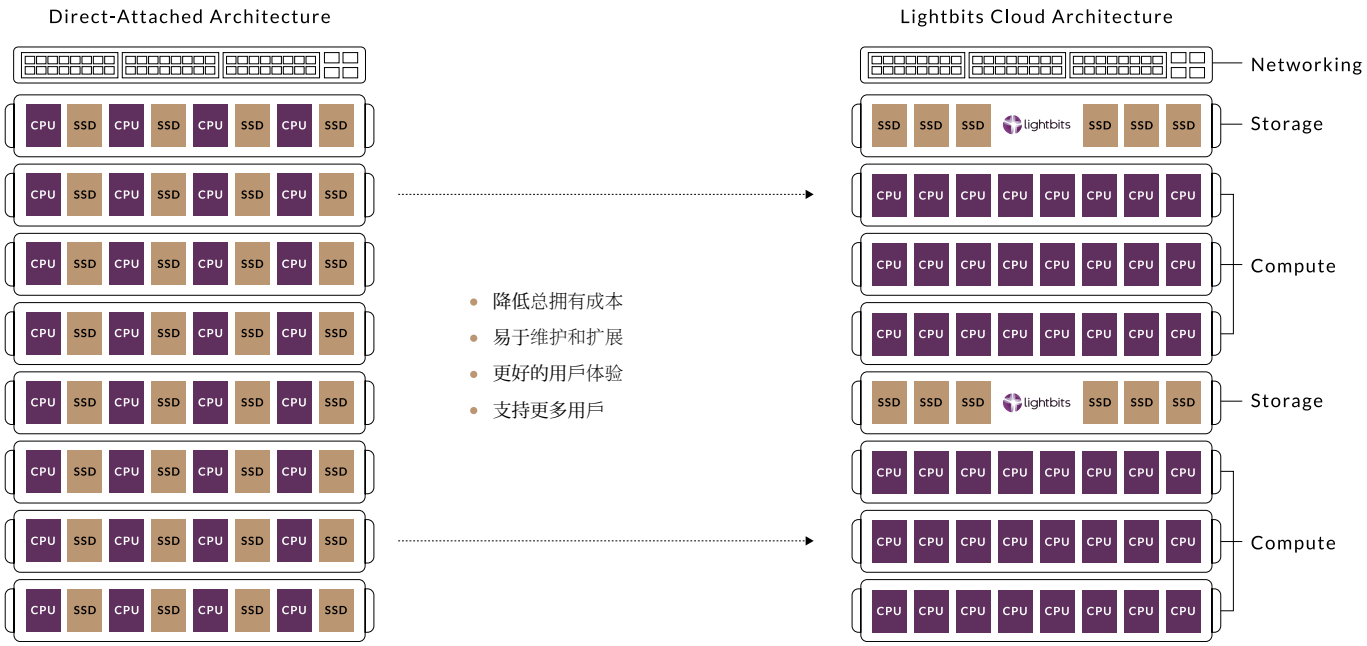

Figure 2: Moving from Direct Attached Storage (DAS) to disaggregated storage and compute

4. Lightbit s Labs 在数据中心部署 NVMe/TCP

Lightbits Labs 解决方案 评估了 节省和性能优势:

- 与直接连接存储 (相比，尾部延迟减少了 50%
- SSD 容量利用率翻了一倍
- 数据服务的性能提高了 2 - 4 倍
- 扩展到数万个节点
- 支持数百万个 IOPS，平均延迟小于 200µs

Lightbits 解决方案在不影响系统稳定性或安全性的情况下实现了这些改进

- 应用服务器及其存储的物理分离
  - 允许独立部署、可伸缩性和升级
  - 使存储基础设施比计算基础设施的扩展速度更快
  - 提高应用服务器和存储的效率
  - 通过应用服务器和存储硬件的独立生命周期管理，简化管理并降低
- 提供 可与内部 NVMe SSD 的 高性能和低延迟，
- 利用现有的网络基础设施相媲美，无需更改
- 支持 多跳 点 m ul ti hop 数据中心网络体系结构

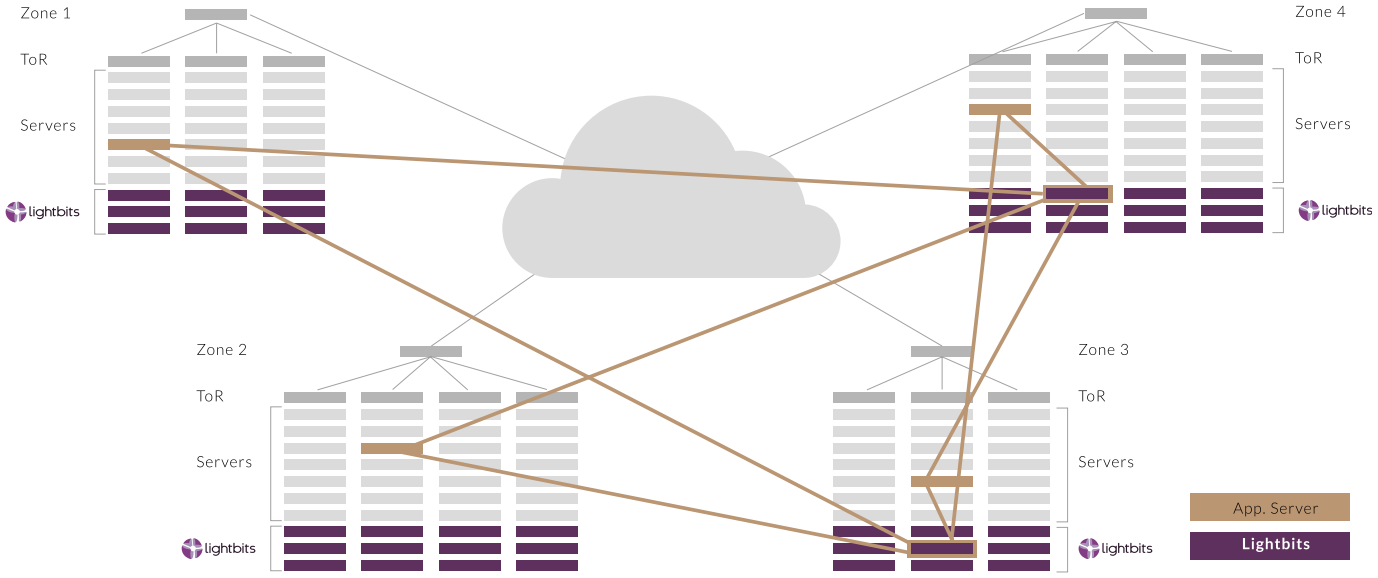

图 3: NVMe / TCP 将存储节点连接到整个数据中心的应用程序服务器

5. Lightbits 存储解决方案的工作原理

Lightbits Lab s 为云和数据中心基础设施提供了一个分 离 的 闪存存储 平台。云规模的网络暴露了当数以万计的计算节点将直接连接的存储孤岛锁定在每个物理节点中时所存在的极端复杂性。 Lightbits 的解决方案 展现了 一个 分离 的高性能 SSD 解决方案的潜力。它使数据中心能够从低效的直接附加的 SSD 模型转移到一个共享模型，在该模型中，计算和存储是独立的，以最大限度地提高资源利用率和灵活性。

在Lightbits Labs 发明 NVMe/TCP 时，我们继续使用用于 DAS 设备的 NVMe 模型，然后将其映射到行业标准的 TCP/IP 协议套件。 NVMe/TCP 将多个并行 NVMe I/ O 队列映射到多个并行 TCP/IP 连接。 NVME 和 TCP 之 间的配对产生了一个简单的、基于标准的、从端到端的并行体系结构。

Lightbits Labs利用现代云架构的并行性实现NVMe / TCP

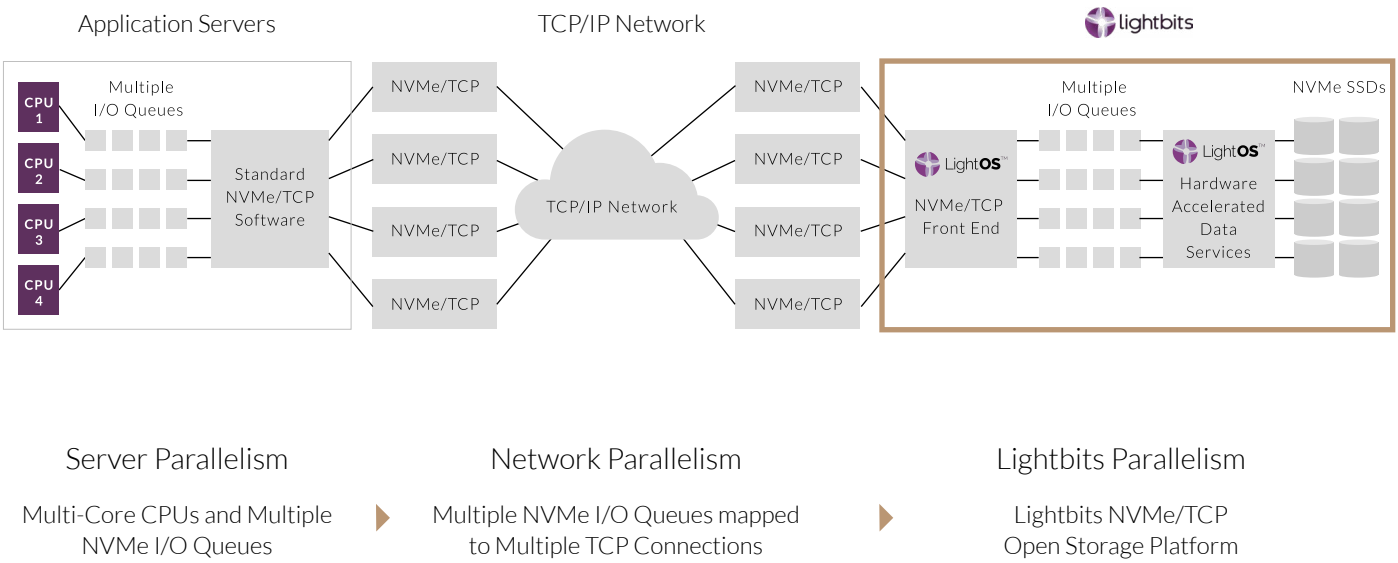

图4：为并行云架构构建的 NVME/TCP

这种新的共享模型使用了创新的NVMe/TCP 标准，它不会因延迟而做折中，也不需要更改网络基础设施或应用服务器软件。 Lightbit s Labs 正在与其他市场领导厂商合作，以采用这一新的 NVM e /TCP 标准。使用Lightbit s L abs 分离存储解决方案，存储通过瘦配置提供给应用服务器。瘦配置意味着管理员可以将任意大小的卷分配给客户端。

并且，只有当应用服务器写入数据时，才消耗底层存储容量。因此，存储在最后可能的时刻才被使用，即当它被需要的时候。这进一步降低了成本，推迟了存储资源的购买。 Light bits 还为线速度的数据服务提供了一种硬件加速解决方案。

因此，当使用 Lightbit s 薄配置技术和硬件加速数据服务时，存储成本可以降低到只相当于 DAS 解决方案成本的一小部分。

5.1. 闪存友好写入算法

Flash媒体的读写延迟都很低。但是，SSD 上的闪存控制器必须持续执行 垃圾收集 操作，以便为传入的写入提供空闲空间。与写入可以覆盖现有数据的硬盘驱动器不同，闪存驱动器只允许将数据写入以前未写入或删除的闪存块。

垃圾收集操作会导致写入放大。顾名思义，SSD 控制器执行垃圾收集时，应用服务器发出的单个写可以放大为更多的写入到实际的闪存媒体上。写入放大增加了在闪存驱动器上的磨损，这减少了它的长期使用。 此外，背景垃圾收集会增加传入 I/O 的延迟，并且随着更多随机写入到闪存驱动器，垃圾收集急剧增加。不幸的是，很高比 例的 I/OS 是随机的。

总的来说，这意味着用户没有获得最好的性能或闪存耐久性。 Lightbit s 实验室解决方案通过一个智能管理层来解决这个问题，该层管理不同服务质量 (级别的 SSD 池。该解决方案减少了 SSD 后台操作，使 I/OS 更快、更高效。该结构将多种算法紧密结合在一起，以达到优化性能和 FLASH 利用率的目的。这包括将数据保护算法与数据服务的硬件加速解决方案紧密耦合，以及我们的高性能读写算法。最后，所有 IO 都在 SSD 池之间进行管理和平衡，从而极大地提高了闪存利用率。该设计提高了总体性能，减少了尾部延迟，减少 了 SSD 上的写入放大和磨损。这意味着 Light OS 为您的闪存提供了最大的投资回报 。

5.2. 高性能数据保护方案

将存储从应用服务器中分离出来，需要智能和高效的数据保护，不影响性能。 L ightbits 结合了高性能的数据保护方案，与数据服务和读写算法的硬件加速解决方案一起工作。 对于 将数据写入 SSD 池 方面 ，与传统的 RAID 算法相比， Lightbits 的 数据保护方法防止了使 SSD 遭受更大磨损的过多写入。

6. 总结

Lightbits Labs 实现了高效的闪存分解，在实现和操作上具有以下优点：

- 不需要任何昂贵的专用网络硬件 轻量级的解决方案运行在标准的 TCP/IP 网络上。
- 使用 TCP/IP 在局域网上或多个局域网上运行，不受协议限制。
- 提供与 DAS 类似的性能和延迟，包括比 DAS 延迟更少 50% 的尾延迟。
- 将高性能数据保护方案与其硬件加速的数据服务解决方案结合起来，以及确保性能不受损害的读写算法。
- 利用硬件加速的解决方案，最大限度地提高闪存效率，使数据服务以全线速度运行，不影响性能。
- 实现供应不足的存储量，从而实现 随增长付费 的消 费模式。

Lightbits 是 NVMe/TCP 的发明者，也是采用它的 推动者 。 作为 一种 新概念的应用 Lightbits的 NVMe/TCP 解决方案 可以实现 有效的闪存分解， 从而获得与 DAS 同样的或 比 DAS 更好的性能。 Lightbit s 已经创建了一个现代的 IP 存储体系结构实现，它最大限度地利 用了应用服务器、 NVMe 、 TCP 和 SSD 并行架构。 Lightbits Labs 使得 云 属性 应用可以实现云规模的性能， 减少了 云数据中心 的 TCO

联系我们获取更多信息: info@lightbitslabs.com

About 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.lightbitslabs.com info@lightbitslabs.com

US Office:

1830 The Alameda,
San Jose, CA 95126, USA

Israel (Kfar Saba) Office:

17 Atir Yeda Street,
Kfar Saba, Israel 4464313

Israel (Haifa) Office:

3 Habankim Street,
Haifa, Israel 3326115