# 概述

NVMe传输是一种抽象协议层，旨在提供可靠的NVMe命令和数据传输。为了支持数据中心的网络存储，通过NVMe over Fabric实现NVMe标准在PCIe总线上的扩展，以此来挑战SCSI在SAN中的统治地位。

# 协议

## FC

NVMe-oF的目标是未来存储Fabric互连的霸主，想替代现有的Fibre Channel（光纤通道）、iSCSI、FCoE（实际上没普及就淡出了）三种SAN块存储协议。

## RDMA

NVMe over Fabric支持把NVMe映射到多个Fabrics传输选项，主要包括FC、InfiniBand、RoCE v2、iWARP和TCP。

然而，在这些Fabrics选项协议中，我们常常认为InfiniBand、RoCE v2(可路由的RoCE)、iWARP是理想的Fabric，其原因在于它们支持RDMA。

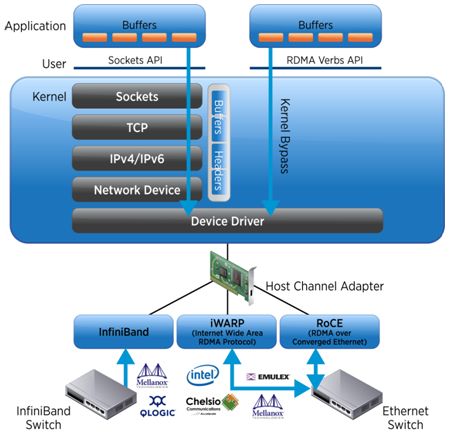
InfiniBand(IB)：从一开始就支持RDMA的新一代网络协议。由于这是一种新的网络技术，因此需要支持该技术的网卡和交换机。

RDMA融合以太网(RoCE)：一种允许通过以太网进行RDMA的网络协议。其较低的网络头是以太网头，其上网络头(包括数据)是InfiniBand头。这允许在标准以太网基础架构(交换机)上使用RDMA。只有NIC应该是特殊的，并支持RoCE。

互联网广域RDMA协议(iWARP)：允许通过TCP执行RDMA的网络协议。在IB和RoCE中存在功能，iWARP不支持这些功能。这允许在标准以太网基础架构(交换机)上使用RDMA。只有NIC应该是特殊的，并支持iWARP(如果使用CPU卸载)，否则所有iWARP堆栈都可以在SW中实现，并且丢失了大部分的RDMA性能优势。

那么为什么支持RDMA在选择NVMe over Fabric时就具有先天优势？这要从RDMA的功能和优势说起。

RDMA是一种新的内存访问技术，RDMA让计算机可以直接存取其他计算机的内存，而不需要经过处理器耗时的处理。RDMA将数据从一个系统快速移动到远程系统存储器中，而不对操作系统造成任何影响。RDMA技术的原理及其与TCP/IP架构的对比如下图所示。



因此，RDMA可以简单理解为利用相关的硬件和网络技术，服务器1的网卡可以直接读写服务器2的内存，最终达到高带宽、低延迟和低资源利用率的效果。如下图所示，应用程序不需要参与数据传输过程，只需要指定内存读写地址，开启传输并等待传输完成即可。RDMA的主要优势总结如下：

1) Zero-Copy：数据不需要在网络协议栈的各个层之间来回拷贝，这缩短了数据流路径。

2) Kernel-Bypass：应用直接操作设备接口，不再经过系统调用切换到内核态，没有内核切换开销。

3) None-CPU：数据传输无须CPU参与，完全由网卡搞定，无需再做发包收包中断处理，不耗费CPU资源。

这么多优势总结起来就是提高处理效率，减低时延。那如果其他网络Fabric可以通过类似RD6MA的技术满足NVMe over Fabric的效率和时延等要求，是否也可以作为NVMe overFabric的Fabric呢？下面再看看NVMe-oF和NVMe的区别。

NVMe-oF和NVMe之间的主要区别是传输命令的机制。NVMe通过外围组件互连Express(PCIe)接口协议将请求和响应映射到主机中的共享内存。NVMe-oF使用基于消息的模型通过网络在主机和目标存储设备之间发送请求和响应。

NVMe-oF替代PCIe来扩展NVMe主机和NVMe存储子系统进行通信的距离。与使用本地主机的PCIe 总线的NVMe存储设备的延迟相比，NVMe-oF的最初设计目标是在通过合适的网络结构连接的NVMe主机和NVMe存储目标之间添加不超过10 微秒的延迟。

## TCP

# DBDK

# SPDK

# 发展未来

我们将NVMe替代SCSI和SAS作为SSD设备的默认连接。高端部署将使用NVMe over Fabrics，在现有的阵列平台中使用NVMe将会保留了快照，复制，压缩和重复数据删除等功能，并采用了像Excelero和Apeiron这样的新平台体系结构。过去，这些功能的缺乏使得基于NVMe的产品没有受到关注。然而，随着时间的推移，NVMe肯定会取代传统全闪存技术中存在的传统架构。