白皮书

2014 年 10 月

数据中心中的 RoCE

简介

在数据不断增长的环境中，所有数据的快速传输对于高效使用信息至关重要。基于远程直接内存访问 (RDMA) 的互连可为提升数据中心效率、降低整体复杂性以及提高数据交付性能提供理想选择。RDMA 使数据可以从存储传输到服务器，而无需通过 TCP/IP 以太网的 CPU 和主内存路径传递数据。可获得更高 CPU 和整体系统效率，因为存储和服务器的计算能力只用于计算（而不是处理）网络流量。

RDMA 可实现亚微秒级延迟和高达 56Gb/s 的带宽，从而可转化为令人惊叹的快速应用程序性能、更好的存储和数据中心利用率以及简化的网络管理。

不过直到最近，只在 InfiniBand 结构中提供了 RDMA。随着基于融合以太网的 RDMA (RoCE) 的出现，基于以太网或混合协议结构的数据中心也可利用 RDMA 的优势。

背景

由于存在一些非常基础的挑战，因此基于 TCP/IP 或 UDP 的传统互连协议的效率远远低于 RDMA。

首先，这些传统协议需要在数据传输的发送和接收端都将数据写入内存缓冲区。这会使宝贵的资源无法用于 CPU 的主要计算职责，而是改为将它们专用于输入/输出进程对内存缓冲区进行的重复复制和读取。

此外，套接字 (socket) API 用作应用程序访问网络的接口，这需要双向通信。必须先传递发送请求，并且必须收到针对请求的响应确认和授予权限，然后才能开始传输。互连过程中的这一额外步骤会增加整体传输时间，并且会耗尽远程设备上的计算资源。

另一方面，RDMA 旨在应对这些挑战。通过将操作系统旁路、零复制和 CPU 卸载内置在体系架构中，对 RDMA 进行规划以获得高性能。

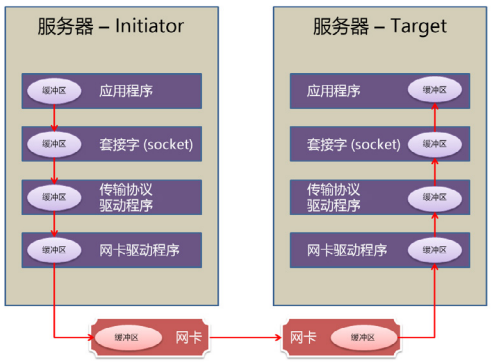

图 1. *TCP/IP 通信*

操作系统旁路使应用程序可以直接访问网卡，从而使 CPU 可以直接与 I/O 适配器通信，无需操作系统从用户空间转换到内核。借助 RDMA，无需操作系统或驱动程序参与，从而可大量节省互连事务的效率。

通过 RDMA 还可在无需将数据复制到内存缓冲区的情况下进行通信。这种零复制传输使接收节点可以直接从发送节点的内存读取数据，从而减少 CPU 参与所形成的开销。

而且与传统互连不同，RDMA 提供了可由硬件处理的传输协议栈。通过从软件卸载栈，CPU 参与程度更低，并且传输更加可靠。

RDMA 通过操作系统旁路、零复制和 CPU 卸载实现显著降低 CPU 开销是为了最大程度提高效率，从而提供闪电般的快速互连。

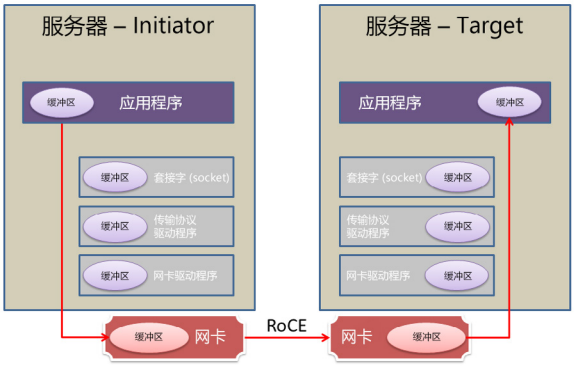

图 2. *RDMA 通信*

RDMA 情况

当前的数据中心需要基础互连在极低延迟的情况下提供最高带宽。无论市场如何，低延迟都已成为绝对必要的因素。

例如，移动、游戏和视频点播领域使用低延迟来确保实时的一致响应。在需要高性能计算的金融市场中，极低延迟可能意味着数百万美元的差别。数据中心的横向扩展需要更高性能，存储到服务器和服务器到存储事务也是如此。同样，迁移到固态磁盘 (SSD) 存储也会使延迟与存储市场相关。

尽管高带宽十分重要，但是如果没有低延迟，带宽就没什么价值。通过网络移动大量数据可以使用 TCP/IP 来实现，但是只有 RDMA 才能实现低延迟，避免成本高昂的传输延迟。而且，RDMA 卸载可减少抖动，这意味着低响应时间的一致性显著提高。

为何选择 RoCE？

CIO 和应用程序编写者很早就认识到了 RDMA 的优势，因此倡导使用 InfiniBand 基础架构。尽管如此，某些 IT 经理不愿意从现有以太网数据中心迁移或学习新协议。

基于融合以太网的 RDMA (RoCE) 可实现 RDMA 的所有优势，不过是在现有以太网网络上实现。借助 RoCE，无需将数据中心从以太网转换为 InfiniBand，从而使公司可节省大量资本支出。对于应用程序而言，使用基于 InfiniBand 的 RDMA 与使用基于以太网的 RDMA 之间没什么区别，因此 RoCE 可很好地适用于更熟悉以太网环境的应用程序编写者。

基本上，RoCE 最终将 RDMA 技术引入基于以太网的数据中心，从而使这类数据中心可以受益于 RDMA 的低延迟，而不必采用基于 InfiniBand 的网络基础架构。

RoCEv2

最新版本的 RoCE 甚至添加了更出色的功能。通过更改数据包封装以包含 IP 和 UDP 标头，现在可以跨二层和三层网络使用 RDMA。这可实现第 3 层路由，从而将 RDMA 引入具有多个子网的网络。得益于更新的版本，IP 多播现在也成为可能。

结论

在 RoCE 出现之前，对于解决糟糕的数据中心性能，只有两个非常有限的选择。但是借助 RoCE，在性能和节省方面出现了一个极佳的选择。RoCE 可在现有以太网基础架构上实现高效数据传输，从而可提供 InfiniBand 的许多优势，而无需花费支出来添加大量硬件或进行大规模转换。

由于 RoCE 的出现，最终可以在传统以太网数据中心中体验最低互连延迟。

北京迈络思科技有限公司

咨询电话：+86-10-57892000

销售咨询：china_sales@mellanox.com

市场合作：marketing_cn@mellanox.com

*欲了解更多欢迎登陆www.mellanox.com

微信

15-3951WP
修订版 1.1