DOI: 10.19392/j.cnki.1671-7341.202018105

RDMA 网络传输技术研究综述

金 浩[1] 杨洪章[2]

1.南京中兴新软件有限责任公司 江苏南京 210012;2.中兴通讯股份有限公司 广东深圳 518057

摘 要:面对高性能计算、分布式存储等应用的快速发展,现有的网络件架构无法满足业务发展的需求,新兴的 RDMA 技术解决了传统网络架构的性能瓶颈,本文对 RDMA 技术的基本原理深入研究,并为应用开发给出指导。

关键词:RDMA; RoCE; 零拷贝

RDMA(Remote-Direct Memory Access) 远程内存直接访问,由 Infiniband 公司针对高性能技术领域推出的高速网络技术,与传统网络技术相比,RDMA 能够提供更高带宽、更低时延、占用更少的系统资源。

1 RDMA 关键技术

RDMA 基本原理是本地应用通过网卡直接访问远端节点的内存数据,无需远端 CPU 和操作系统的参与,主要包含下面几种关键技术:第一,内核旁路技术,应用程序直接使用 RDMA 接口实现数据发送、接收,不需要使用系统调用,避免了在系统态、用户态之间切换的开销。第二,减少拷贝,RDMA 网卡能够直接访问主机内存空间,将上层应用设计为访问固定物理内存空间,可以实现全流程零拷贝。第三,减少资源占用,RDMA 网卡与主机内存之间采用 DMA(Direct Memory Access) 方式,占用系统总线,不占用 CPU 资源,因而报文收发流程 CPU 开销很小。

2 RDMA 实现方式

RDMA 规范的前身 Infiniband 简称 IB,起初用于高性能计算领域,需要使用专用的交换机、路由器等网络设备,部署维护成本高。为了降低 RDMA 使用成本,推动 RDMA 技术普及,业界厂家将 IB 协议移植到以太网协议上,定义了 RoCE(RDMA over Converged Ethernet) 、iWarp(internet Wide Area RDMA protocol) 两种协议。RoCE 分为 v1、v2 两种,v1 基于 Ethernet 协议实现 IB 协议,不支持跨网络传输; v2 基于 UDP 协议实现 IB 协议,支持三层路由设备,适合大规模组网。iWarp 则是在 TCP 协议之上实现 iWarp 协议,对网络设备要求低,但性能较差,目前只有 intel 生产支持 iWarp 协议的网卡。

3 RDMA 通信原理

RDMA 协议定义 RC、UC、UD 三种通信模式。RC(Reliable Connection) 模式,保证报文正确的传输到目的端,支持报文 ack 确认、超时重传,某个报文超时没有确认,则重传该报文后的所有报文。UC(Unreliable Connection) 模式,需要提前建链,报文不需要携带地址信息,不支持 ack 确认、重传,不保证对端能正确接收。UD(Unreliable Datagram) 模式,不需要建链,每个报文都携带目标地址、目标队列信息,不支持 ack 确认、重传,每个报文不能大于网络 MTU 限制。三种模式稳定性依次下降,执行效率依次升高,RC、UC 链路资源都需要占用网卡的 cache 资源,并发链路数量过多时,需要考虑 UD 模式。

协议定义了双边、单边 2 种通信原语。send、recv 指令属于双边原语,接收端执行 recv 指令等待数据到达,发送端执行 send 指令发起数据传输,双边 CPU 都参与传输过程,适合小数据传输。read、write 指令属于单边原语,得知远端内存地址后,本地网卡直接访问远端内存,远端 CPU 无感知。单边原语是 RDMA 规范中最具创新性的特性,通过 RDMA 协议把本地内存总线延伸到其他主机,传输效率高,适合较大数据的传输。不同模式下支持原语不同,RC 模式支持全部原语,UC 模式不支持 read,UD 模式仅支持 send、recv 单边操作。

4 部署及应用

常用的 RDMA 网卡硬件分为 IB 卡和 RoCE 卡两类。RoCE 网卡兼容传统以太网卡操作,支持 ip、ifconfig 等系统命令管理网卡设备,上层应用可以同时访问以太网卡、RoCE 网卡,通用 socket 接口访问以太网卡,专用 RDMA 接口访问 RoCE 网卡。IB 网卡则需要安装专用的驱动软件,并与 IB 交换机连接,部署成本高。IB 网络通过 ID 信息访问网卡,为了使 IB 网卡与传统网卡兼容,驱动软件提供了 ipoib 模块、opensm 服务,前者支持 tcp/ip 与 IB 协议互转,实现通过 IP 地址访问本地 IB 网卡,后者周期扫描子网内所有 IB 网卡,并维护网络路由信息,实现通过 IP 地址与其他 IB 节点通信的能力。

网卡硬件对上层应用透明,应用程序通过专门的接口库实现 RDMA 通信,不需要区分不同硬件。socket 应用也可以通过 ipoib 模块直接运行到 RDMA 网络上,该方式使用内核协议栈,无法发挥 RDMA 性能,从测试结果看,ipoib 方式的传输带宽下降 37%,时延增加 376%,CPU 资源开销则增加 86~490%。

5 性能测试

为验证 RDMA 传输性能,本文采用 Mellanox ConnectX-3 Pro 网卡对 RDMA、TCP 的传输性能做对比测试。网卡配置为 RoCE 模式,qperf 测试 TCP 性能,ib_send_bw、ib_send_lat 测试 RoCE 性能,分别测试 64B、256B、1KB、4KB、16KB、64KB 几种大小数据包。测试结果显示,64 字节小包时,RDMA 带宽达超过 TCP 的 10 倍,平均时延不足 TCP 的 4%; 64K 字节大包时,RDMA 带宽超过 TCP 带宽 73%,平均时延只有 TCP 的 7%; RDMA 传输的 CPU 开销远小于 TCP。本性能测试充分证明,同等硬件的 RDMA 性能远远优于传统 TCP 链路。

6 总结

与传统网络相比,RDMA 在带宽、时延、资源占用方面优势显著,目前 RDMA 在很多顶级产品都已商用,如阿里的 PolarFS、华为的 FusionStore、亚马逊服务集群。可以预见,不久的将来,将会看到更多的厂家推广应用 RDMA 技术。

参考文献:

[1] 吴昊,陈康.基于 RDMA 和 NVM 的大数据系统一致性协议研究[J].大数据,2019.04: 89-99.

[2] 陈游旻,陆游游.基于 RDMA 的分布式存储系统研究综述[J].计算机发展与应用,2019.02: 227-239.

基金课题:国家重点研发计划项目(2018YFB1003302) ; 江苏省工业和信息产业转型升级专项资金项目; 南京市工业和信息化发展专项资金项目

作者简介:金浩,男,高级工程师,主要研究方向为高速存储及网络协议栈。