DOI:10.19353/j.cnki.dzsj.2021.21.073

RDMA高性能通信库的设计与实现

苏州高等职业技术学校　石宏华

伴随着信息化、数据化的时代现状，电子信息通信架构的技术是否安全可靠，将成为计算数据处理中心面临的挑战之一，同样也成为当前热门的研究方向之一。由于限制数据集群处理技术多半是通信技术的因素，所以在实际研究阶段可从该方面入手，而RDMA（远程数据直接读取技术）可优化与完善传统网络架构的网络传输带宽类问题。鉴于此，本文依据远程数据直接读取技术为基础，探究其在高性能通信库中的作用与价值。

1 RDMA高性能通信技术

1.1 无限带宽技术架构

何为无限带宽技术？正所谓，无限带宽是由英文InfiniBand直译过来，其作为一种广泛应用于高性能计算的计算机网络通信标准之一，可以代替传统的PCI总线技术，并支持全双工工作方式，以此也具备极高的吞吐量和极低的延迟等优势。InfiniBand架构简称为IB架构，在高性能通信库设计阶段，即需要相关硬件的支持，同样也要在软件协议层下开展设计方案。故此，在设计阶段，须按照IB架构制定软件层面的修改方案，并同时进行操作系统环节的软硬件相互配合，达到实现最大化通信的基本目的。需注意一点的是：IB架构设计的目的在于应对服务器之间的带宽问题，其协议栈如图1所示。

1.2 RDMA技术

RDMA（远程数据直接读取技术）是一种类似于直接存储器存取，但是与其不同的是本地直接存储器存取须经过CPU单元模块，而RDMA则不需要通过上述的单元模块，直接完成数据的传输过程，以此也体现了其优质的数据传输效率。基于RDMA研发出的零拷贝技术，已经在当前高性能数据集群中得到较好的应用。

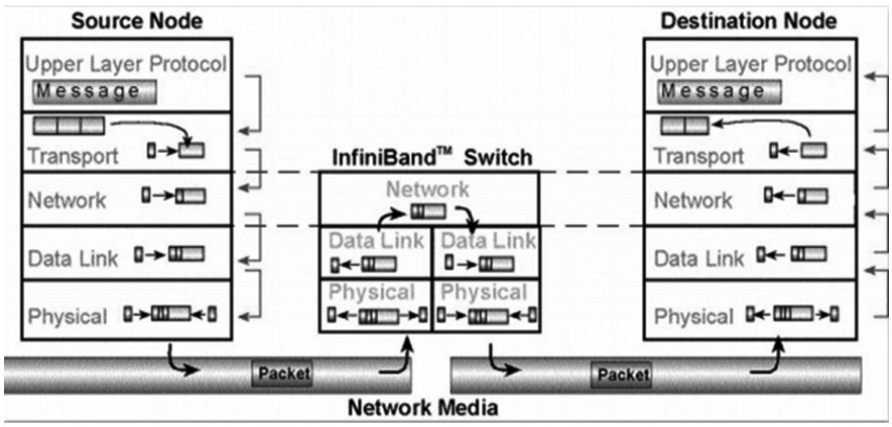

图1 IB架构协议栈

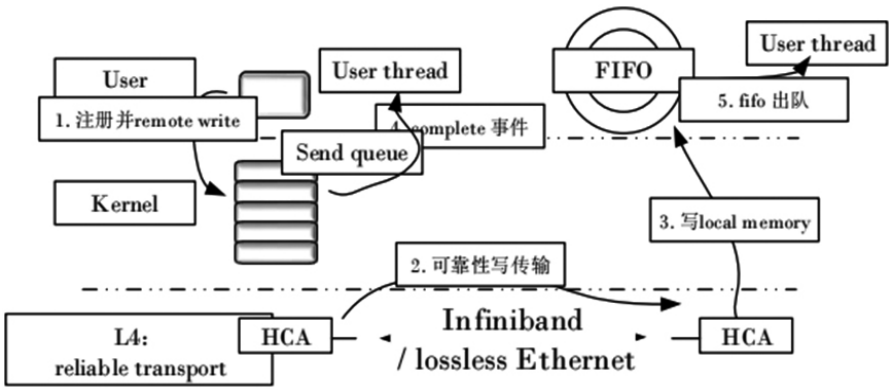

图2 IBverbs通信方案设计图

2 RDMA高性能通信方案的设计

2.1 基本设计流程

IB设计框架下实现RDMA可完成直接收发数据的基本功能，而其中便涉及到简单数据接口，即IBverbs。在设计阶段，可充分利用数据接口IBverbs提供的现有接口，并借助远程数据直接读取技术的额外特点，制定出最佳的设计方案。比如，可在IBverbs接口上设计类TCP的连接方案，以便实现计算机集群运算通信的基本要求。在IBverbs接口实现数据信息发送传输阶段，须严格按照以下操作流程进行。一是，完整注册对端直接方案的内存块；二是，将数据发送消息加入到队列中，此处是利用IBverbs自带的send接口模块；三是，利用send单元模块的设计程序即可完整一次数据的发送。此外，信息数据的接收，同样会遵循严格的设计流程，其具体的通信设计流程见图2所示。

2.2 多线程设计

为了保证数据传输的可靠新与安全性，通常选择多线程模块化设计方案，即利用多个线程池完成数据的优化处理，其中各个线程之间均保持独立，不会额外占用其他线程的硬件资源，以保证数据传输的高效性。对外阶段，须根据FIFO的设计功能，保证流水的基本需求，以此防止因无流水功能而出现的带宽利用率不高的情况。采用发送队列与接收队列好处还在于减少数据传输的时间。在FIFO可从两个角度入手，即可以设计成实现队列的阻塞接收方案，可以设计成非阻塞接收方案。此外，基于RDMA环境下，接收端若想要得到数据信息，仅仅通过事前注册好的数据块到相应的驱动模块中，即可完整信息的获取，而不在需要拷贝的方式。须注意一点的是，上层应用接收端会无法预知信息数据的传输到达时间，一旦数据调用接口IBverbs存在问题，将限制数据的传输，故此在RDMA机制下可使用其自身的通信协议栈主动完成数据的管理。鉴于此，通过以上的多线程的简要分析，上层应用调用接口时，须根据着重数据块的大小设计。

综上所述，在RDMA高性能通信库的设计阶段，根据RDMA在基本设计流程与多线程设计方面的分析，现选择IBverbs接口模块为基础，并结合多线程设计模式，完成RDMA技术支持的高性能阻塞通信接口的设计环境。

3 RDMA高性能通信库的实现

3.1 RDMA通信库框架的设计

上文中得出以IB架构下的高性能通信网卡，此为物理层；IBverbs接口模块则主要是担负数据的发送与接收。RDMA通信库框架设计，包含多个层次，如以内存注册与内存映射的设计为例。在RDMA技术中内存映射方式是指，用户在调用内存注册接口时会产生并获取虚拟地址，系统中则会根据产生的虚拟地址创建到与之对应的物理地址中。简言之，IB架构中的表保存机制便是实现物理地址与虚拟地址的对接，经过实际的测试发现，RDMA内存注册维持的时间过长，而降低了信息传输的性能。为解决此弊端，可使用预注册Buddy算法管理机制。该算法机制，可将内存也合并成大小不同的内存块，且每一个内存块内部均有标志位标识，以此保证了通信数据库的大数据块传输。故此，利用Buddy系统管理之后的内存注册，极大程度上方便了高性能通信库的管理与应用。

3.2 RDMA通信库性能分析

对于高性能通信库的设计，多半集中在集群间点对点通信环节中，实现带宽利用率最大化的阻塞式高性能通信接口。鉴于此，后续的性能分析阶段，则更多的侧重于带宽情况与GPU支持情况。下面将简要介绍一下实现高性能通信库之后的性能调节与优化情况。

（1）高性能通信调优的基本原则

高性能通信库架构在设计完成之后，侧重于一个角度的观察，即带宽的利用效率。一个角度可从两个基本原则上入手，一是，认真作好系统程序检查工作，保证设计过程中没有任何编程错误，同时也须检查计算机设备支持的性能优化模式；二是，从本质上进行提升通信库的性能，如优化内存注册方面、应用Buddy注册算法机制等。以上两点均可改善通信库的基本性能。

（2）零拷贝对延迟的影响

要想探索零拷贝对延迟带来的影响，可侧重于数据传输时间延时，并做出简要设计，即拷贝延迟+协议处理延迟+网络传输时间。其中在网络传输时间方面，仅研究应用层感知的数据延迟环节即可。经过一系列的计算分析得出，零拷贝引入了较大的网络传输影响，而针对于不同大小的数据块传输时呈现正相关关系。同时，在计算中发现，RDMA的延迟效果低于其他类，凸显其自身的低延迟特性。也说明了RDMA的优势可体现在大数据块的传输阶段。

（3）高性能通信库测试标准

本次研究设计的高性能通信库侧重于在一些上层应用提供阻塞发送数据接口，故此在实际的测试评估中直接构造数据发送模型即可。从高性能通信库的性能效益模块入手，在带宽较高的网络拓扑结构中体现的较为显著，故此研究重点须放到带宽的最大有效率环节中。此外，对于通信库数据延迟到达问题，同样是在日后测试环节中须重视的一个角度。以系统测试为例，IB架构组网环境下，首先，制定数据注册时间的测试，其次，完成不同数据块传输的差异性对比，以此获得点对点数据传输的延迟系数与带宽系数，务必保证每一个环节的操作规范性。故此不再作出的赘述。

总结：综上所述，面对海量的数据资源，高性能集群技术的发展已经迫在眉睫，而实现集群技术与高速通信的相互融合，便成为行业亟待解决的问题。本文基于RDMA技术下设计出高性能通信库解决方案，并提出Buddy算法等一系列技术，为上层应用的可扩展性提供了依据，同时经过测试得出该设计方案符合带宽最大利用率的需求。

作者简介：石宏华，男，硕士，现供职于苏州高等职业技术学校，助讲，研究方向：电子信息。