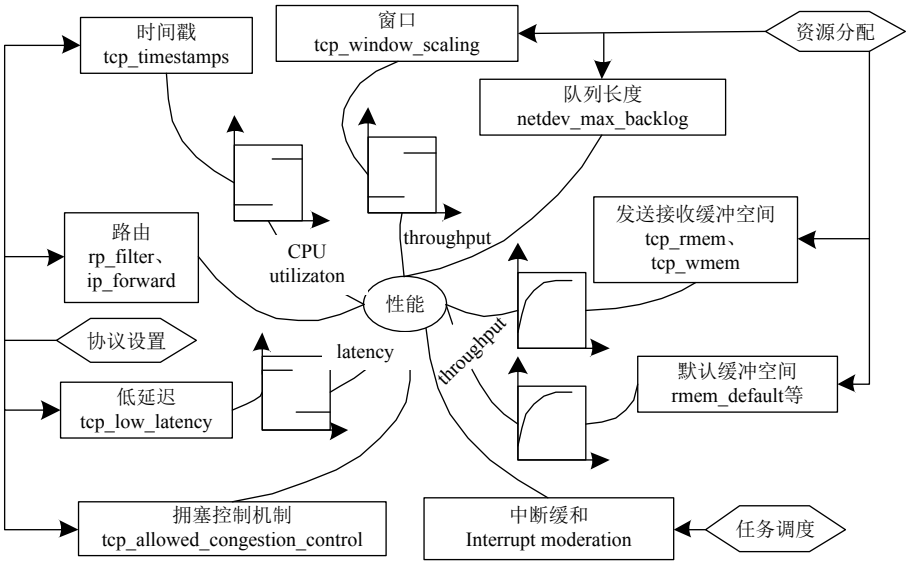

图 5 网络参数优化模型图

4.4 避免应答延迟 (Delayed ACK)

TCP 采用应答延迟机制时, 如果当前时间与最近一次接收数据包的时间间隔小于延迟应答超时时间, 则会推迟 ACK 应答的发送, 积攒多个应答并将它们结合成一个响应包, 与需要沿该方向发送的数据一起发送, 从而减少协议开销. 然而, 在应用程序进行交互处理时, 延迟 ACK 应答时间过长可能会降低应用程序的效率. TCP 协议中利用宏定义 TCP_DELACK_MIN 控制最小延迟确认时间, 一般默认值为 (HZ/25), 也就是 40 ms. 本文在不改变其它参数的情况下, 逐一试验在 1 ms~40 ms 范围内不同的 TCP_DELACK_MIN 值, 并测试网络最大带宽, 发现最小延迟应答时间设为 5 ms 左右时, 网络带宽可以达到最大, 既能维持较低协议开销, 又可以减少 TCP 传输中 ACK 的等待时间, 使得网络带宽最大化, 提升内存的利用率. 最佳的延迟应答时间受服务器系统环境和网络应用场景的影响, 本文所得结果在其它集群系统中可能不是最优, 但其试验方法具有普适性.

5 实验结果

本文在国产异构众核系统上对 IPoIB 进行测试, 配备 32 GB 内存, 节点间采用 40 GB/s 的 Infiniband EDR 网络连接. 网络性能测试工具选用 Netperf-2.4.5 和 Iperf-2.0.2. Netperf 主要用于记录两对联节点间 TCP 单连接带宽、延迟、CPU 利用率、内存等资源的占用情况, Iperf 用于记录两对联节点之间 TCP 多连接带宽.

5.1 优化效果分析

5.1.1 不同消息大小带宽对比

利用 Netperf 工具测试两对联节点间单个 TCP 连接优化前后不同消息大小带来的带宽变化, 结果如图 6 所示.

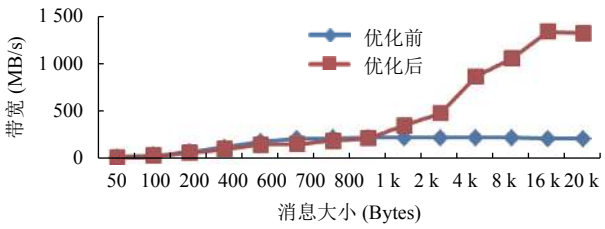

图 6 不同消息大小带宽对比图

从图 6 可以看出, IPoIB 通信性能优化效果良好, 优化后的 IPoIB 网络带宽明显高于优化前. 由测试数据可知, 优化后的峰值带宽达到 1340 MB/s, 对比优化前的 227 MB/s 提升近 6 倍. 可见针对 IPoIB 的优化对带宽具有较好提升效果, 使得系统的 IB 网络资源得到尽可能的使用, 提升了 IPoIB 在国产异构并行系统上的运行效果, 证明优化方法足够有效.

5.1.2 多连接带宽对比

使用 Iperf 测试两节点间多个 TCP 连接的网络最大带宽, 从 2 个连接测到 10 个连接, 如图 7 所示.

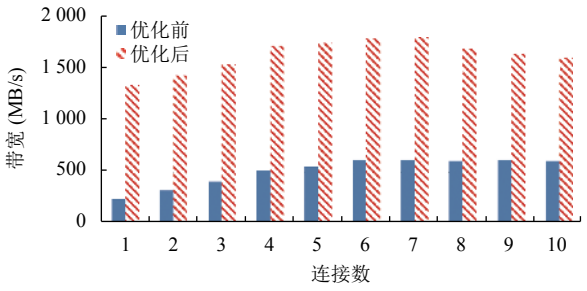

图 7 不同连接数带宽对比图

测试数据表明, 连接数为 6 或 7 时, 网络带宽达到最大. 优化后的网络带宽可以达到 1800 MB/s, 相比单连接最大带宽 1340 MB/s 提升 1.34 倍, 相比优化前多连接的网络最大带宽 615 MB/s, 提高近 3 倍. 可以看出, 在多个进程下网络带宽能有较大提升, 主要原因是: (1) 并行系统的 CPU 多核组得到了有效利用; (2) 通过在传输过程中对数据包进行抓取, 发现在其中某个 TCP 连接等待 ACK 不发送数据包时, 另一个 TCP 连接不用等待 ACK, 继续发包, 从而使得链路带宽得到有效利用.

5.1.3 CPU 利用率对比

在高速网络环境下, CPU 的处理能力很大程度上影响网络的性能. 通过测试发现, 优化前 IPoIB CPU 利用率为 33%, 优化后的 IPoIB CPU 利用率在 24% 左右, 优化后的 CPU 利用率明显低于优化前. 在最好的情况下, IPoIB 优化后的 CPU 利用率可以降低约 30%.

5.2 与万兆以太网的性能对比

利用 Netperf 测试优化后的 IPoIB 在国产并行系统上两节点间通信的带宽和延迟, 并与万兆以太网对比. 万兆以太网卡型号 T520, 10 GB/s, 理论带宽为 1250 MB/s, 万兆网卡测试环境采用国产中标麒麟服务器, 处理器型号为申威 1621. 测试结果如表 1 所示.

表 1 IPoIB 与万兆以太网性能对比

测试项	带宽 (MB/s)	延迟 (μs)
IPoIB	1340	39
万兆以太网	1024	30

从表 1 可以看出万兆以太网的稳定带宽在 1024 MB/s 左右, 而 IPoIB 优化后的带宽达 1340 MB/s, 是万兆以太网持续带宽的 1.31 倍; 万兆以太网延迟平均约为 30 μs, IPoIB 延迟平均 39 μs, 相比万兆以太网延迟高 9 μs. 基础性能对比结果表明, IPoIB 在国产异构众核并行系统上的通信性能相比基于 10 GbE 的万兆以太网通信性能要略显优势, 持续带宽优于万兆以太网, 延迟虽然比万兆以太网略大, 但实际应用中聚合带宽是主要考虑因素, 因此 IPoIB 相比 10 GbE 是更好的选择.

5.3 乱序处理效果分析

利用 Iperf 测试国产并行系统上两节点间通信处于不同流量负载时的乱序情况, 乱序处理窗口长度 W 设为 32, 为消除偶然因素影响, 每个负载下运行 10 次取平均值. 表 2 显示了在不同流量负载下乱序处理前后每秒发生的乱序次数并统计乱序减少比例. 可以看出, 乱序处理对于减少乱序数据包的作用效果十分明显: 网络流量较高时乱序情况比较严重, 而经过乱序处理后乱序次数明显减少, 乱序减少比例可达 95% 以上; 当网络负载较轻时, 经过乱序处理后网络不再有乱序包.

表 2 不同流量负载下乱序处理效果

发送速率 (MB/s)	乱序处理前后乱序次数		乱序减少比例 (%)
	处理前 (次/s)	处理后 (次/s)	
1800	3815	184	95.2
800	1411	37	97.4
400	469	0	100

为了验证乱序较重时窗口长度 W 设置过大对性能造成的不利影响, 每隔一段时间让发送方故意丢弃一次数据包以模拟乱序较重的情况, 利用 Iperf 测试窗口长度分别为 32 和 80 的最大网络带宽, 结果如表 3 所示. 从表 3 可以看出, 乱序较重时, W 为 32 的最大带宽为 1782 MB/s, 乱序程度减少 95.8%; W 为 80 的最大带宽为 1624 MB/s, 即便乱序减少效果更好, 但带宽下降明显. 实验结果证明窗口长度不宜设置过大, 过大反而会造成带宽性能下降.

表 3 不同窗口长度的网络性能对比

窗口长度	乱序减少比例 (%)	最大网络带宽 (MB/s)
32	95.8	1782.32
80	98.3	1624.71

6 结束语

本文将 IPoIB 移植到国产异构众核并行系统上, 并进行了乱序处理、拷贝优化、网络参数调优以及应答延迟避免等一系列优化措施. 测试结果显示, 优化后 IPoIB 基础带宽峰值性能为 1340 MB/s, 比优化前 IPoIB 带宽提升近 6 倍, 也高于 10 GB 万兆以太网; 多连接下带宽达到 1800 MB/s, 相比单连接提升 1.34 倍;

CPU 利用率也有了显著降低; 乱序处理机制作用效果明显.

IPoIB 基于 IB 的 send/receive 异步消息机制实现, 而没有利用具有零拷贝、CPU 负载卸载优势的 RDMA 机制, 考虑到在一些特定的应用场景下利用 RDMA 实现 IPoIB 的通信效果可能会更好, 后续将制定以 RDMA 为底层通信机制的 IPoIB 实现策略, 以期进一步提高 IPoIB 通信性能.

参考文献

1 徐迪威, 余焯佳. InfiniBand 高速互连网络设计的研究. 电脑与电信, 2012, (7): 26–29. [doi: 10.3969/j.issn.1008-6609.2012.07.025]

2 刘爱华, 钱德沛, 董小社, 等. IPoIB 体系结构及其应用. 计算机科学, 2003, 30(9): 85–88. [doi: 10.3969/j.issn.1002-137X.2003.09.025]

3 朱叶青, 牛德姣, 蔡涛, 等. 不同网络环境下大数据系统的测试与分析. 江苏大学学报 (自然科学版), 2016, 37(4): 429–437. [doi: 10.3969/j.issn.1671-7775.2016.04.010]

4 何王全, 刘勇, 方燕飞, 等. 面向国产异构众核系统的 Parallel C 语言设计与实现. 软件学报, 2017, 28(4): 764–785. [doi: 10.13328/j.cnki.jos.005197]

5 Dalessandro D, Devulapalli A, Wyckoff P. Design and implementation of the IWARP protocol in software. Proceedings of International Conference on Parallel and Distributed Computing Systems. Phoenix, AZ, USA. 2005. 471–476.

6 Kaur G, Kumar M, Bala M. Comparing Ethernet and soft RoCE for MPI communication. IOSR Journal of Computer Engineering, 2014, 16(4): 52–58.

7 秦宣龙, 李大刚, 都政, 等. 面向数据中心网络的高速数据传输技术. 软件, 2016, 37(9): 1–7. [doi: 10.3969/j.issn.1003-6970.2016.09.001]

8 伍卫国, 杜哲君, 刘娟, 等. InfiniBand 结构中 SDP 协议分析. 微电子学与计算机, 2004, 21(9): 144–148. [doi: 10.3969/j.issn.1000-7180.2004.09.041]

9 Wright GR, Stevens WR. TCP/IP 详解卷 2: 实现. 陆雪莹, 蒋慧, 译. 北京: 机械工业出版社, 2000. 680–803.

10 孔金生, 任平英. TCP 网络拥塞控制研究. 计算机技术与发展, 2014, 24(1): 43–46.

11 王敏杰, 徐昌彪, 刘光明. 无线网络下 TCP 重传定时器研究. 计算机工程与应用, 2004, 40(36): 146–150. [doi: 10.3321/j.issn:1002-8331.2004.36.046]

12 胡晓峰, 孙志刚, 苏金树. 基于 NewReno 拥塞控制机制的 TCP 分组乱序影响分析. 计算机工程与科学, 2009, 31(5): 8–12. [doi: 10.3969/j.issn.1007-130X.2009.05.003]