超算集群无损RoCEv2网络性能评测

龙汀汀[1]，付振新[1,2]，李若淼[1,2]，龚翔宇[3]，吴涛[3]，樊春[1,2]
（1.北京大学计算中心，北京 100871；2.北京大学长沙计算与数字经济研究院，湖南 长沙 410000；
3.华为技术有限公司，江苏 南京 210012）

摘　要：为了评估无损RoCEv2网络技术在高性能计算（HPC）领域的实际表现，以无损RoCEv2、TCP/IP和InfiniBand这3种网络为测试对象，搭建HPC集群，使用主流的HPC Benchmark和科学计算应用对上述3种网络进行对比测试，获取各网络的基本性能数据以及在科学计算应用场景下的实际表现，还测试了基于RoCEv2的240节点集群的HPL效率。实验结果表明，在超算集群中，无损RoCEv2与InfiniBand有基本相当的性能，且都显著优于TCP网络。随着集群节点数量的增加，RoCEv2网络具有较好的线性可扩展性。无损RoCEv2网络相对于InfiniBand，在保持成本优势的同时具有大致相当的性能。

关键词：高性能计算；RoCEv2；InfiniBand；无损网络
中图分类号：TP393
文献标志码：A
DOI：10.11959/j.issn.1000–436x.2024244

Performance evaluation of lossless RoCEv2 network in HPC cluster

LONG Tingting[1], FU Zhenxin[1,2], LI Ruomiao[1,2], GONG Xiangyu[3], WU Tao[3], FAN Chun[1,2]
1. Computer Center, Peking University, Beijing 100871, China
2. Changsha Institute for Computing and Digital Economy, Peking University, Changsha 410000, China
3. Huawei Technologies Co., Ltd., Nanjing 210012, China

Abstract: To evaluate the practical performance of lossless RoCEv2 network technology in the HPC field, an HPC cluster was set up using three types of networks: lossless RoCEv2, TCP/IP, and InfiniBand. Mainstream HPC benchmarks and scientific computing applications were used to compare these networks, obtaining basic performance data and actual performance in scientific computing scenarios. The HPL efficiency of a 240-node cluster based on RoCEv2 was also tested. The experimental results show that in supercomputing clusters, lossless RoCEv2 has performance roughly equivalent to InfiniBand and is significantly superior to TCP networks. As the number of cluster nodes increases, the RoCEv2 network demonstrates good linear scalability. Compared to InfiniBand, the lossless RoCEv2 network maintains a cost advantage while offering approximately equivalent performance.

Keywords: high performance computing, RoCEv2, InfiniBand, lossless network

0　引言

高性能计算（HPC, high performance computing）是一种利用并行处理技术，整合大量计算资源来处理复杂计算任务的计算方式，被广泛应用于大数据、科学计算、气象气候等领域[1-2]。与多处理器系统相比，高性能计算使用网络将多个服务器节点相连，大幅降低了成本。与此同时，慢速、高时延的传统TCP/IP网络也成为集群的潜在性能瓶

收稿日期：2024–10–24
通信作者：付振新，fuzhenxin@pku.edu.cn
基金项目：湖南创新型省份建设专项经费资助项目（No.2023GK1010）
Foundation Item：Special Funds for Construction of Innovative Provinces in Hunan Province (No.2023GK1010)

颈，所以引入了InfiniBand（IB）[3]、RoCEv1[4]、RoCEv2[5]等高速网络。这些高速网络能否支撑各类高性能计算场景，充分发挥并行计算的优势，同时在性能和成本之间取得平衡，成为高性能计算领域持续关注的问题。故而有必要跟进高速网络技术的发展，评估这些技术对高性能计算的影响。

基于传统TCP/IP架构的网络在传输过程中涉及数据在进程间的多次拷贝，导致传输时延较大、CPU负载较高。针对这一问题，先后提出了InfiniBand、RoCEv1（RDMA over converged Ethernet）、RoCEv2等基于远程直接内存访问（RDMA, remote direct memory access）技术的网络。其中，与RoCEv1相比，RoCEv2使用IP和UDP协议代替InfiniBand网络层协议，从而可以通过传统IP路由器来转发RoCEv2报文，扩展了RoCEv2的适用范围。但UDP缺少可靠传输机制，导致发生丢包时依赖于上层应用进行重传，大大降低了传输效率。因此，各厂商发展不丢包的无损网络技术，不断缩小RoCEv2与InfiniBand之间的性能差距[6-7]。华为近期提出了智能无损网络[8-11]，引入了一系列优化算法，实现整网零丢包和低时延。

为此，本文以华为公司最新的RoCEv2网络为研究对象，以TCP/IP和InfiniBand为对比，搭建高性能计算集群，使用主流的高性能计算基准测试工具和科学计算应用对上述3种网络进行多方面的性能评测。相比于前人在HPC领域所开展的测试工作[6,12-14]，本文的贡献有以下几点。

1) 针对华为近期提出的智能无损RoCEv2网络，使用主流基准测试工具和科学计算应用开展多方面的测试，评估新技术对HPC的影响。

2) 与TCP/IP、InfiniBand网络进行横向对比，分析不同网络的性能，为HPC集群的搭建和升级提供参考。

3) 展望HPC领域中网络的未来发展方向，包括将网络拓扑感知能力引入资源管理系统中以提高任务运算效率、探索新的网络拓扑等。

1　HPC集群中的网络

1.1　RDMA技术

在高性能计算等对网络I/O性能要求极高的场景中，传统TCP/IP架构已难以满足应用需求。这主要是因为TCP/IP需要经过操作系统内核处理，不仅占用大量CPU资源和内存带宽，还需要在系统内存、CPU缓存和网卡缓存之间多次复制数据。这些操作不仅消耗了宝贵的计算资源，还显著增加了通信时延。相比之下，RDMA技术通过将传输协议直接实现在网卡硬件上，采用零复制、内核旁路和CPU卸载等创法，实现了直接访问远程服务器内存的能力，无须操作系统内核参与。这种设计极大地提高了网络通信的吞吐量，同时大幅降低了时延，特别适合在大规模并行计算集群中应用。TCP/IP和RDMA的技术原理对比如图1所示。

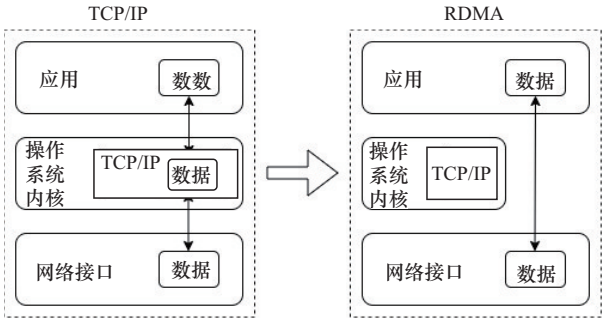

图1　TCP/IP和RDMA的技术原理对比

RDMA的特点如下。

1) 零复制。零复制的网络技术可以实现网卡与应用内存之间数据的直接互相访问，不再需要数据在应用内存和操作系统内核之间多次复制，因此，传输时延会显著减少。

2) 内核旁路。内核旁路技术可以使网卡直接与应用程序内存进行相互的数据访问，而不再需要经过内核，免去了在内核态与用户态之间做环境切换。

3) CPU卸载。RDMA将数据移动的工作从CPU卸载到网卡上。传统的网络通信需要CPU参与数据的拷贝、协议处理和中断处理等操作，这会占用大量的CPU资源。通过CPU卸载，RDMA允许数据直接在主机的内存和网络设备之间传输，避免了CPU对数据路径的干预。这不仅降低了CPU的负载，也减少了内存带宽的消耗，从而提高了系统的整体性能和效率。

常见的RDMA网络包括InfiniBand、RoCE和iWARP。

InfiniBand是一种专门为RDMA设计的高性能网络，具有高吞吐、低时延和高可靠的特点。它被

用于计算机之间的数据互连，可以直接将数据从一台计算机上的存储设备传输到另一台计算机上的用户空间，绕过并避免系统调用的开销。InfiniBand 采用了完全独立的硬件设计和传输协议，如图 2 所示[15]。该协议需要支持专门的 InfiniBand 的交换机和网卡，技术先进，但是成本高昂。

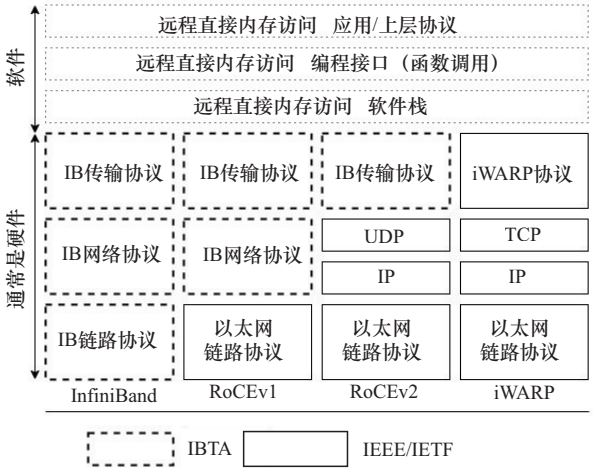

图 2　典型 RDMA 网络协议

RoCE 是由 IBTA（InfiniBand trade association）标准化组织定义的一种基于以太网络的 RDMA 网络协议。它有 2 个版本：RoCEv1 和 RoCEv2。RoCEv1 直接在以太网层上运行，不支持 IP 路由，限制了其在大型网络中的应用。RoCEv2 则在 UDP/IP 层上实现，支持 IP 路由，使其可以在更广泛的网络环境中使用。由于 RoCEv2 网络采用 InfiniBand 网络的传输层，因此它既具有 InfiniBand 网络的零拷贝、低时延、低 CPU 利用率等特点，又能够很好地兼容于以太网络。使用 RoCE 网络时，可用普通的以太交换机配合支持 RoCE 协议的网卡。

iWARP 构建在标准的 TCP/IP 之上，利用现有的 TCP 网络基础设施，无须特殊的网络配置，适用于广域网环境，并且依赖于 TCP 的流量控制和拥塞控制机制。缺点是开销更大，性能弱于 RoCE 网络。该协议可以使用普通的以太网交换机，但是需要支持 iWARP 的网卡。

1.2　华为智能无损网络

在无损网络领域，业界较为成熟的方案包括 IEEE 802.1 中的 PFC 流控技术和采用 ECN 的 DCQCN 算法。华为在此基础上引入优化算法和在网计算技术进行改进，提出了华为智能无损网络。

1.2.1　无损网络优化

华为智能无损网络采用 RoCEv2 协议，在链路层基于 PFC（priority-based flow control）[16]优先级流控机制。在网络层采用强化学习、监督学习和动态启发式探索机制动态优化 ECN（explicit congestion notification）拥塞参数，应对网络实时流量变化、精准反压，从而大幅提高了交换机的缓存利用效率。针对大规模组网环境，创新性地提出了主动网络死锁避免技术，根据拓扑特征加路由策略，计算出破除死锁的最佳路径点，破除网络死锁，实现大规模组网的无死锁，实现整网高可靠。

1.2.2　在网计算技术

针对超算集群中常用的集合通信场景，华为提出了在网计算技术，将集合通信中的部分计算卸载到网络中，减少消息入网的次数，降低时延，提升通信效率，缩短计算任务的完成时间。

一个典型的集合通信场景中，每轮迭代计算参与其中的服务器在计算完成后将结果通过网络传递到主服务器上做聚合计算，之后主服务器再将聚合计算的结果通过网络传递给所有服务器作为下一轮迭代计算的输入数据。如图 3 所示，期间计算发生在服务器侧，数据会在网络中多次传递，而网络只作为转发。

在网计算技术中，在每一个 Leaf 交换机将其连接的每个服务器的计算结果收集并进行一次聚合计算，而后将所有的 Leaf 交换机的计算结果汇聚到一个 Spine 交换机上，由该 Spine 交换机再做聚合计算，最后将得到的结果通过网络分发到各个服务器中，进入下一轮的迭代计算。在网计算技术方案将集合通信的计算卸载到网络中，服务器侧只需要一次数据的收发，数据在网络中传递的过程中完成计算，大大减少了消息进入网络的次数，提升了通信效率，缩短了整个计算任务的完成时间。

2　实验设计

2.1　HPC Benchmark

HPC Benchmark 是一类用于评测高性能计算集群的浮点运算效率、网络性能等关键基础性能指标的测试工具。本文选用的工具包括 HPL（high performance linear system package）和 OSU MPI Benchmark。

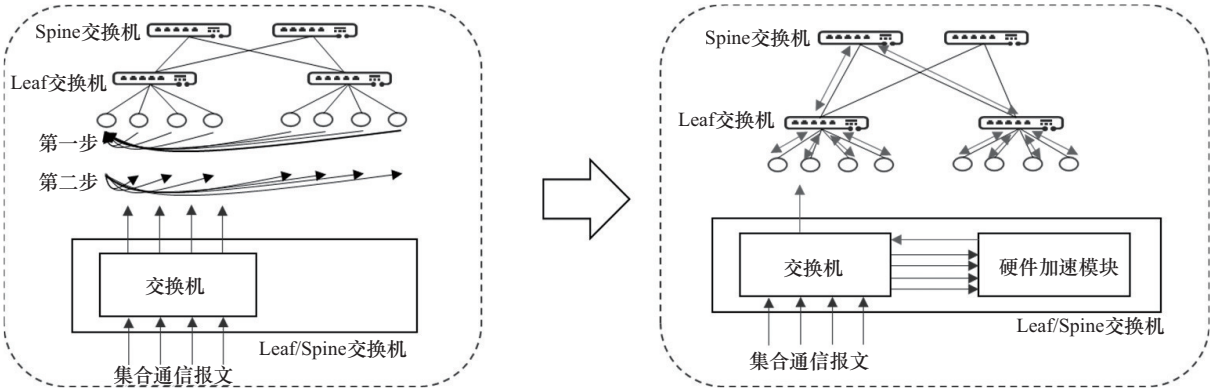

图3　在网计算算法

2.1.1　HPL

HPL是高性能计算领域中广泛使用的基准测试工具，用于评估集群在执行浮点运算时的性能表现[17]，可以充分利用计算机的并行计算能力和高效的通信机制，测试结果以每秒执行的浮点运算次数（FLOPS）来表示。

2.1.2　OSU MPI Benchmark

OSU MPI Benchmark是由美国俄亥俄州立大学提供的一套业界广泛使用的MPI通信效率评测工具[18]，程序生成不同规模的数据并执行各种不同模式的MPI通信，测试包含MPI点对点通信、MPI集合通信等不同通信形式下的网络带宽与时延。

2.2　科学计算应用测试

本文选取在北京大学校级高性能计算平台使用较多的2个应用，分子动力学软件VASP以及地球系统模式CESM，来验证不同网络在真实应用场景下的性能表现。

CESM（Community Earth System Model）[19]是美国国家大气中心开发的地球系统模式，是目前应用最广泛的地球系统模式之一。该程序能够对大气（ATM）、陆面（LND）、陆冰（GLC）、海洋（OCN）、海冰（ICE）、海浪（WAV）、河川径流（ROF）等进行气候模拟。本文基于CESM1.2.2进行实验，利用上述提到的7个模块进行了200天的地球运行模拟，并利用其内置耦合器（CPL）进行模块间的数据交互。

VASP（Vienna Ab-initio Simulation Package）[20]是一款广泛应用于材料科学和计算物理领域的第一性原理计算软件包，支持对晶体、表面、分子等多种体系进行结构优化、能带计算、分子动力学模拟等，为研究材料的微观性质和性能提供了强有力的工具。本文基于VASP5.4.4进行实验。

2.3　实验环境

本文使用了8台硬件配置和软件版本完全一致的服务器作为计算节点。为了避免环境温度等无关变量影响实验结果，关闭了所有节点的超线程和睿频。计算节点软硬件环境如表1所示。

表1　计算节点软硬件环境

型号	Dell PowerEdge R620
CPU	2×Intel® Xeon® CPU E5-2697 v2 @ 2.70 GHz
内存	128 GB DDR3
硬盘	2×320 GB SSD RAID1
网卡	MCX455A
操作系统	CentOS 7.9
网卡驱动	4.9-2.2.4
编译器	icc (ICC) 2021.3.0
MPI版本	Hyper MPI b007, Intel® MPI 2021.3.0
测试软件版本	OSU MPI Benchmark 5.6.2 HPL 2.3 (Intel® Distribution) CESM 1.2.2 VASP 5.4.4

本文分别使用100 Gbit/s TCP/IP、100 Gbit/s RoCEv2和100 Gbit/s InfiniBand网络将以上8台节点互联以进行对比试验。所使用的网络设备如表2所示。其中，测试TCP时关闭了无损，因此本文使用的是标准TCP网络。网络拓扑如图4~图6所示。

为了进一步测试RoCEv2网络在更大规模的集群中的性能，另外搭建了一个240节点的集群，节点型号为Huawei FusionServer Pro X6000 V6，CPU为2×Intel Xeon Platinum 8358，内存容量为256 GB，所使用的网络为RoCEv2。

表 2　网络设备

网络类型	交换机型号	交换机软件版本	厂商
TCP	CE8850-64CQ-EI	CE8850V200R019C10SPC030T	华为
RoCEv2	CE8850-64CQ-EI	CE8850V200R019C10SPC030T	华为
InfiniBand	MSB7790	0.9454	迈络思

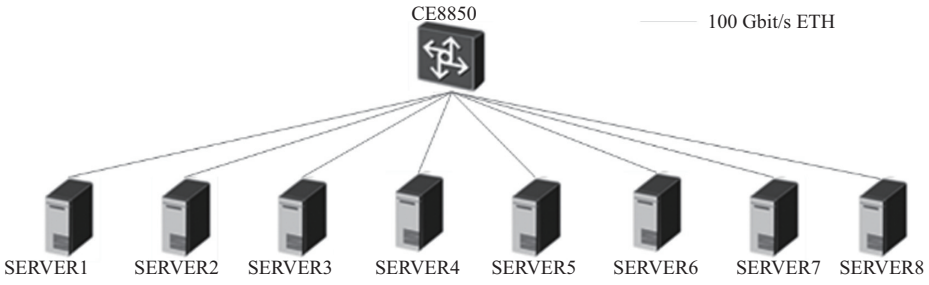

图 4　TCP

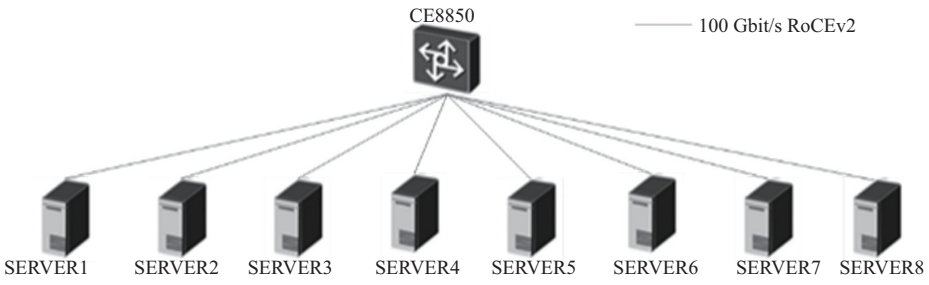

图 5　RoCEv2

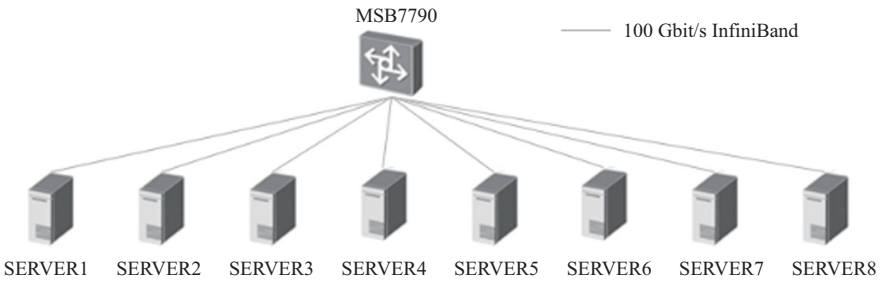

图 6　InfiniBand

3　实验结果与分析

3.1　HPC Benchmark 结果与分析

3.1.1　网络通信吞吐与时延

本文使用 Hyper MPI，运行 OSU MPI Benchmark 工具测量集群中网络通信的性能指标。相关结果如下。

1) MPI 点到点通信吞吐测试。本文采用 OSU MPI Benchmark 分别对 InfiniBand 与 RoCEv2 组网进行 MPI 点到点通信吞吐测试 osu_bibw。从图 7 可以看出，在 100 Gbit/s 带宽下，随着消息长度的增加，InfiniBand 与 RoCEv2 吞吐性能基本相当。

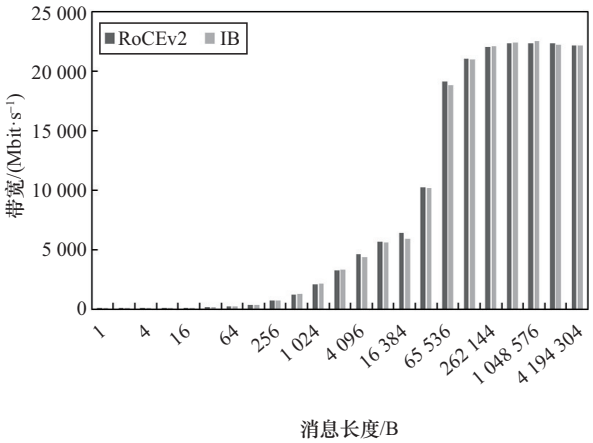

图 7　MPI 点到点通信吞吐测试

2) MPI点到点通信时延测试。本文采用OSU MPI Benchmark分别对InfiniBand与RoCEv2组网进行MPI点到点通信时延测试osu_latency。从图8可以看出，在100 Gbit/s带宽以及不同消息长度下，均体现出了InfiniBand网络转发芯片对比RoCEv2使用的以太网络转发芯片有0.6~0.7 μs的静态时延优势。

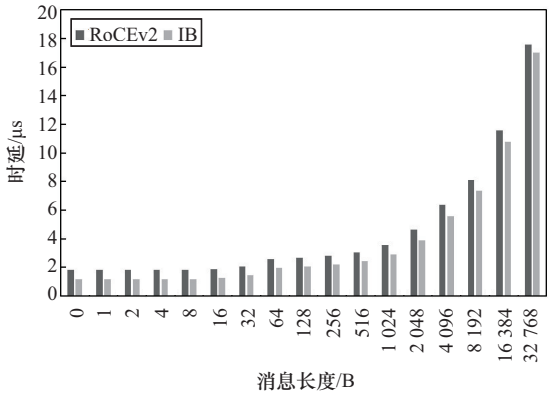

图8　MPI点到点通信时延测试

3) MPI_Allreduce测试。本文采用OSU MPI Benchmark分别对InfiniBand与RoCEv2组网进行MPI_Allreduce测试。从图9可以看出，在单层组网8台服务器100 Gbit/s接入带宽条件下，随着消息长度的增加，InfiniBand与RoCEv2性能基本持平，RoCEv2平均值比InfiniBand优1.9%。

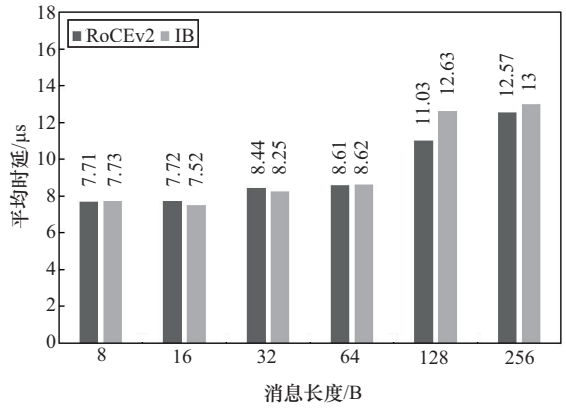

图9　MPI_Allreduce测试

4) MPI_ Alltoall测试。本文采用OSU MPI Benchmark分别对InfiniBand与RoCEv2组网进行MPI_Alltoall测试，在单层组网8台服务器100 Gbit/s接入带宽条件下，从图10可以看出，随着消息长度的增加，InfiniBand与RoCEv2性能基本持平。

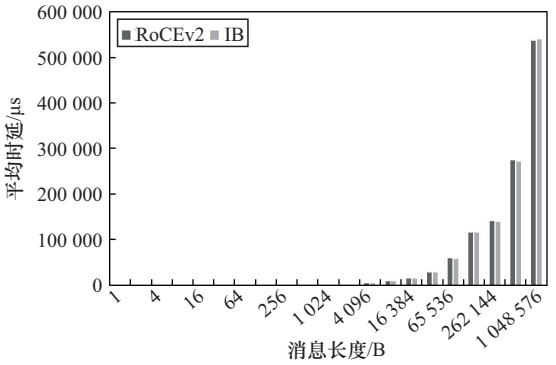

图10　MPI_Alltoall测试

3.1.2　浮点运算性能(HPL)结果

使用Intel® MPI，分别在每个计算节点单独运行HPL，测试结果如表3所示。

表3　单节点HPL测试结果

计算节点	浮点运算性能/GFLOPS	效率
SERVER1	497.503	95.97%
SERVER2	497.333	95.94%
SERVER3	497.600	95.99%
SERVER4	497.666	96.00%
SERVER5	497.700	96.01%
SERVER6	497.539	95.98%
SERVER7	497.647	96.00%
SERVER8	497.657	96.00%

单节点的理论浮点运算性能为518.4 GFLOPS。效率值为实际浮点运算性能除以理论浮点运算性能。由于关闭了睿频，各节点的浮点运算性能相差很小，极差仅0.367 GFLOPS，表现非常稳定，有助于避免处理器性能波动带来的偏差，从而更准确地反映集群的网络性能。8节点HPL测试结果如图11所示。

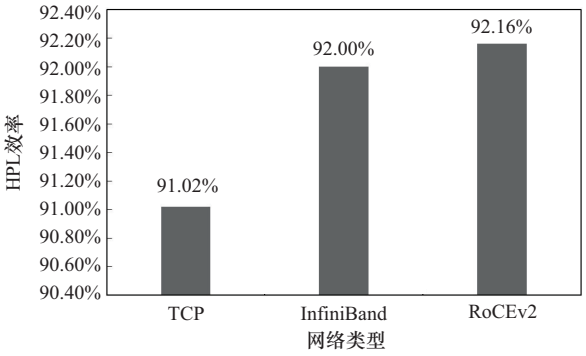

图11　8节点HPL测试结果

3.2　科学计算应用测试

为验证 RoCEv2 在科学计算中的性能表现，本文使用 Intel(R) MPI 在真实应用场景下运行科学计算应用，将 RoCEv2 与 TCP 以及 InfiniBand 进行对比。为保证结果的可靠性，本文设置数据规模以保证程序运行时间超过 1 200 s。实验结果如图 12 和图 13 所示。在 TCP 环境下，VASP 的运行时间远远超过 3000 s 且程序挂起，故本文认为 VASP 不适宜在该交换机下进行 TCP 模式运行。

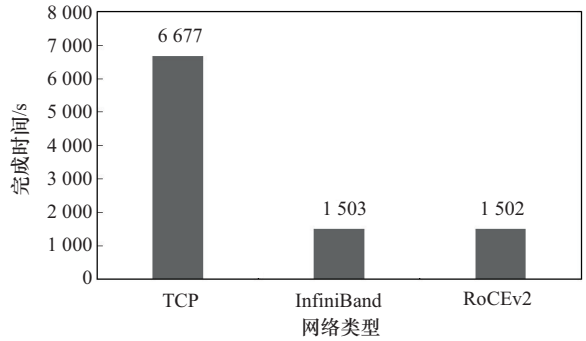

图 12　CESM 8 节点测试结果

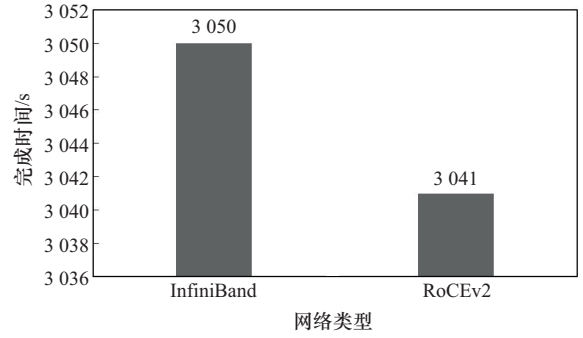

图 13　VASP 8 节点测试结果

实验结果表明，RoCEv2 与 InfiniBand 具有近乎一致的性能表现，这也验证了真实应用场景下的 RoCEv2 对于 InfiniBand 的替代作用。相较 TCP，RoCEv2 与 InfiniBand 运行程序时间更短，这说明 RoCEv2 与 InfiniBand 协议在缩小通信时延上具有很好的表现。

3.3　240 节点 HPL 效率测试

为了验证 RoCEv2 网络在大规模集群中的可扩展性，测试了不同节点数的集群的 HPL 效率，最多使用了 240 节点。测试结果如图 14 所示。测试表明，RoCEv2 网络具有较好的线性可扩展性，其性能随集群规模增长而下降的幅度较小。

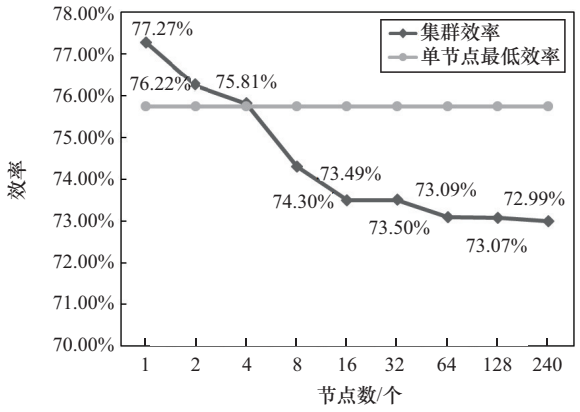

图 14　240 节点 HPL 测试结果

4　结束语

本文搭建并测试了采用智能无损 RoCEv2 网络的高性能计算集群，并将其与 TCP/IP 网络、InfiniBand 网络进行了对比。本文所使用的 RoCEv2 网络使用了智能无损技术，该技术通过减少丢包现象来有效缓解了重传导致的性能下降。该网络相比于广泛采用的 InfiniBand 网络，性能差距较小，且成本更低。本文的测试结果表明在科学计算应用和 Benchmark 测试中，RoCEv2 网络在性能方面的瓶颈并不显著。

从实验结果中可以看出，在超算集群中，智能无损 RoCEv2 与 InfiniBand 有基本相当的性能，且都显著优于 TCP 网络。在 OSU MPI Benchmark 的 4 个测试项目中，RoCEv2 与 InfiniBand 在吞吐和 Alltoall 测试中基本相当，在时延测试中 InfiniBand 有一定优势，而在 Allreduce 中 RoCEv2 则有一定优势。在 HPL 浮点运算性能测试中，RoCEv2 与 InfiniBand 基本相当，且均优于 TCP 网络。在科学计算应用测试中，RoCEv2 和 InfiniBand 基本相当，且均展现了相对于 TCP 的显著性能优势。在 240 节点 HPL 效率测试中，RoCEv2 网络表现出了较好的线性可扩展性，240 节点的 HPL 效率仅从单节点的 77.27% 下降至 72.99%。综合来看，基于华为智能无损网络的 RoCEv2 在保持成本优势的同时，在高性能计算领域具备和 InfiniBand 大致相当的性能。然而，本文所开展的实验也存在一定局限，所选应用及 Bechmark 不够全面，无法覆盖高性能计算的所有使用场景，未来有待利用更大规模的集群和更多场景的计算任务来进一步验证 RoCEv2 网络的性能特点。

展望未来，高性能计算集群的网络发展方向可能包括如下 2 个方面。首先，利用 RoCEv2 网络拓扑信息来优化计算任务调度。随着计算集群规模的扩大，计算任务也将随之增加，为计算任务分配更合适的算力资源可以显著提高集群的运行效率。未来，北京大学和华为将尝试把网络拓扑感知能力引入资源管理系统中，从而提高算力资源分配方案的合理性。此外，近年来面向高性能计算和存储的新型组网拓扑在成为业界的一个重要课题，对系统的性能提升起了重要的作用。例如，Fujitsu 在 Fugaku 系统中采用了 6D-Torus 拓扑、IBM 在 Summit 系统中采用了 Fat-tree 与 Hybrid Cube Mesh 相结合的拓扑[21]。后续将尝试在计算集群中使用更多的新型网络拓扑。

参考文献：

[1] BENACCHIO T, BONAVENTURA L, ALTENBERND M, et al. Resilience and fault tolerance in high-performance computing for numerical weather and climate prediction[J]. The International Journal of High Performance Computing Applications, 2021, 35(4): 285-311.

[2] SCHMIDT B, HILDEBRANDT A. Next-generation sequencing: big data meets high performance computing[J]. Drug Discovery Today, 2017, 22(4): 712-717.

[3] InfiniBand Trade Association. About InfiniBand™[R]. 2021.

[4] KAUR G, BALA M. Rdma over converged ethernet: a review[J]. International Journal of Advances in Engineering & Technology, 2013, 6(4): 1890.

[5] Infiniband Trade Association. InfiniBand trade association releases updated specification for remote direct memory access over converged Ethernet (RoCE) [R]. 2014.

[6] SHPINER A, ZAHAVI E, DAHLEY O, et al. RoCE rocks without PFC: detailed evaluation[C]//Proceedings of the Workshop on Kernel-Bypass Networks. New York: ACM Press, 2017: 25-30.

[7] DECUSATIS C. Handbook of fiber optic data communication[M]. Amsterdam: Elsevier Inc, 2013.

[8] OLMEDILLA C, ESCUDERO-SAHUQUILLO J, GARCIA-GARCIA P J, et al. DVL-lossy: isolating congesting flows to optimize packet dropping in lossy data-center networks[J]. IEEE Micro, 2021, 41(1): 37-44.

[9] GONZALEZ-NAHARRO L, ESCUDERO-SAHUQUILLO J, GARCIA PJ, et al. Efficient dynamic isolation of congestion in lossless datacenter networks[C]//Proceedings of the ACM SIGCOMM 2019 Workshop on Networking for Emerging Applications and Technologies. New York: ACM Press, 2019: 15-21.

[10] GONZALEZ-NAHARRO L, ESCUDERO-SAHUQUILLO J, GARCÍA P J, et al. Modeling traffic workloads in data-center network simulation tools[C]//Proceedings of the 2019 International Conference on High Performance Computing & Simulation (HPCS). Piscataway: IEEE Press, 2019: 1036-1042.

[11] OLMEDILLA C, ESCUDERO-SAHUQUILLO J, GARCÍA P J, et al. Optimizing packet dropping by efficient congesting-flow isolation in lossy data-center networks[C]//Proceedings of the 2020 IEEE Symposium on High-Performance Interconnects (HOTI). Piscataway: IEEE Press, 2020: 47-54.

[12] LI H, CHEN X L, SONG T, et al. Performance of the 25 Gbps/100 Gbps fullmesh RoCE network using mellanox ConnetX-4 lx adapter and Ruijie S6500 Ethernet switch[C]//Advances in Intelligent Systems and Computing. Berlin: Springer, 2019: 757-767.

[13] YU Y, QIN W, TIAN Y J, et al. Performance evaluation of HPC cloud cluster[R]. 2018.

[14] PASE D M. Linpack HPL performance on IBM eServer 326 and xSeries 336 servers[R]. 2005.

[15] iWARP. RDMA here and now technology brief [R]. 2021.

[16] IEEE. 802.1Qbb – priority-based flow control [R]. 2021.

[17] DONGARRA J J, LUSZCZEK P, PETITET A. The LINPACK Benchmark: past, present and future[J]. Concurrency and Computation: Practice and Experience, 2003, 15(9): 803-820.

[18] MVAPICH. MPI over InfiniBand, Omni-Path, Ethernet/iWARP, and RoCE [R]. 2021.

[19] HURRELL J W, HOLLAND M M, GENT P R, et al. The community earth system model: a framework for collaborative research[J]. Bulletin of the American Meteorological Society, 2013, 94(9): 1339-1360.

[20] VASP - Vienna Ab initio Simulation Package. [R]. 2021.

[21] 谭光明, 薛巍, 翟季冬, 等. 2018-2019 中国计算机科学技术发展报告[M]. 北京: 机械工业出版社, 2019.
TAN G M, XUE W, ZHAI J D, et al. 2018-2019 China computer federation proceedings [M]. Beijing: China Machine Press, 2019.

[作者简介]

龙汀汀（1997– ），男，湖南邵阳人，北京大学助理工程师，主要研究方向为高性能计算。

付振新（1995– ），男，天津人，北京大学工程师，主要研究方向为高性能计算。

李若淼（1988– ），男，四川成都人，北京大学高级工程师，主要研究方向为高性能计算。

吴涛（1985– ），男，江苏南京人，华为技术有限公司技术专家，主要研究方向为数据中心网络、网算协同。

龚翔宇（1982– ），男，江苏无锡人，华为技术有限公司技术专家，主要研究方向为数据中心网络。

樊春（1977– ），男，重庆人，北京大学正高级工程师，主要研究方向为高性能计算。