数据中心无损网络关键技术与组网策略研究

Research on Key Technologies and Networking Strategies for Lossless Networks in Data Centers

蔡　毅，樊　蓉，金　沙（中国移动通信集团设计院有限公司上海分公司，上海 200060）
Cai Yi，Fan Rong，Jin Sha（China Mobile Group Design Institute Co.，Ltd. Shanghai Branch，Shanghai 200060，China）

摘　要：
面向算力网络的多元化、多样化、高速化发展趋势以及数据中心网络架构的演进趋势，为解决飞速增长的高性能处理需求、数据存储和算力处理效率问题，在分析总结远端内存直接访问（Remote Direct Memory Access，RDMA）技术的基础上，研究了无损网络的优势应用场景，提出了无损网络组网技术策略，经实测验证，组网性能满足高性能场景需求。

关键词：
算力网络；无损网络；RDMA；组网技术
doi：10.12045/j.issn.1007-3043.2024.07.016
文章编号：1007-3043（2024）07-0083-05
中图分类号：TP393
文献标识码：A
开放科学（资源服务）标识码（OSID）：

Abstract：
In response to the trend of diversified and high-speed development of computing power networks and the evolution trend of data center network architecture，in order to address the rapidly growing demand for high-performance processing，data storage，and computing power processing efficiency，and based on the analysis and summary of RDMA technology，the advantageous application scenarios of lossless networks are studied，and the networking technology strategies of lossless network are proposed. Through actual testing，the network performance is verified to meet the requirements of high performance scenarios

Keywords：
Computing power network；Lossless network；RDMA；Networking technology

引用格式：蔡毅，樊蓉，金. 数据中心无损网络关键技术与组网策略研究[J]. 邮电设计技术，2024（7）：83-87.

0 引言

随着数字中国建设的不断加速，算力网络向多元化、多样化、高速化方向发展。一方面，图形处理器（Graphics Processing Unit，GPU）、现场可编程逻辑门阵列（Field Programmable Gate Array，FPGA）、中央处理器分散处理单元（Data Processing Unit，DPU）等高性能异构计算芯片不断涌现[1]，近5年来，处理器计算性能提升约90倍，同步即时处理数据量增长近百倍；另一方面，全闪存储的出现推动了存储产业升级换代，固态硬盘读写能力较传统机械硬盘提升近百倍，非易失性高速传输总线（Non-Volatile Memory Express，NVMe）的高性能与传统光纤总线（Fibre Channel，FC）存储网络技术发展进度不再匹配。计算、存储对数据中心内网络提出了集中化、低时延、高吞吐、0丢包等要求，需要通过无损网络技术，促进算力、网络、存储

基金项目：上海市2023年度“科技创新行动计划”软科学研究项目（23692123500）
收稿日期：2024-05-17

相互匹配、协同[2]。

现有主流数据中心内网络正逐步由大二层向“IP-CLOS”的组网架构演进,底层基于传统以太网络架构,二层 spine 交换机和 leaf 交换机之间通过 full-mesh 方式连接,即软硬件架构仍主要基于传统传输控制协议/网际协议(Transmission Control Protocol/Internet Protocol,TCP/IP),其与生俱来的技术特征在面向 AI 计算和分布式存储等应用时,出现了高中央处理器(Central Processing Unit,CPU)消耗、存储处理延时达数十微秒、多次内存拷贝、丢包重传等问题[3]。

为解决上述时延和确定性等问题,RDMA 联盟和无限带宽协会(InfiniBand Trade Association,IBTA)主导提出了基于 RDMA 的无损网络技术。RDMA 采用智能网卡和软件架构优化,无需操作系统和 TCP/IP 协议栈介入,以零复制网络技术和内核内存旁路技术实现高性能远程直接数据存取,可将服务期内数据传输时延降低至 1 μs 以下,极大地减轻了 CPU 的负担[4]。

目前,无损网络是数据中心网络演进的新方向,电气与电子工程师协会(Institute of Electrical and Electronics Engineers,IEEE)、中国通信标准化协会(China Communications Standards Association,CCSA)、开放数据中心委员会(Open Data Center Committee,ODCC)等国内外团体均在数据中心无损网络方面开展了标准化研究工作[5]。本文分析 RDMA 无损网络技术特点,研究无损网络的优势应用场景,结合产业生态的成熟度,提出无损网络组网技术策略,并通过实测验证其组网性能,最后探讨无损网络的进一步发展,为无损网络的组网架构和技术演进、提升高性能计算和全闪存储部署效率提供参考,促进无损网络向服务算力网络发展。

1 RDMA 技术特点

1.1 RDMA 技术分类

主流 RDMA 技术的发展经历了无限带宽(InfiniBand,IB)、iWARP、聚合以太网 RDMA(RDMA over Converged Ethernet,RoCE)3 个阶段。作为最早期的 RDMA 技术,IB 提供了一种基于通道的点对点消息队列转发模型[4],无需操作系统和其他协议栈的介入,具备低时延、高可靠、高服务质量的特性,但专有产品与私有协议存在兼容性差的问题,导致运维成本较高。随着以太网的发展,为提供具备高带宽和无丢包能力的网络,iWARP 和 RoCE 技术应运而生。其中,iWARP 技术对底层以太网的构建有极高的要求,因此建设成本也相应较高,这在一定程度上制约了其广泛部署。RoCE 包含 2 个版本:RoCE v1 与 RoCE v2,与 IB 技术相比,仅在网络层和以太网链路层存在差异。RoCE 具备低成本、高兼容性等优势,在以太网中可实现与 IB 相当的时延和带宽性能,具有较高的综合性价比[6]。随着数据中心内网络的演进,目前主流云计算资源池网络大多部署了 RoCE v2 技术,而 RoCE v1 已经很少被使用。

RoCE v2 技术可基于现有主流交换机进行部署,它通过基于优先级的流量控制(Priority-based Flow Control,PFC)以及显式拥塞通知(Explicit Congestion Notification,ECN)来有效提升网络性能。PFC 和 ECN 技术原理如图 1 所示。PFC 用于 2 种流量在以太网中共存时,确保存储流量无丢包且对其他流量无影响;ECN 利用 IP 报文头中的 DS 域来标记报文传输路径上的拥塞状态[7],从而降低发送速度、最终控制拥塞。在网络实际部署中,难点主要在于缓存水线配置调优的复杂度,这种复杂度直接影响业务的性能,在客观上制约了异构组网的实施。

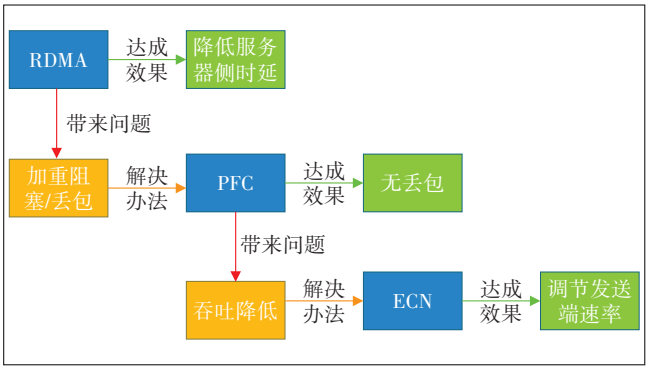

图 1　PFC 和 ECN 技术原理

1.2 技术特点对比

从产业链、网络配置、网络规模等角度,对传统以太网技术和 RDMA 无损网络技术进行对比,具体如表 1 所示。针对数据中心内网络,RoCE 适合应用在高性能场景中,但其在互通性和适用性方面仍需要持续演进和提升。

2 组网技术策略

无损网络技术的开发需要和应用系统深入融合优化,以实现应用系统与计算、存储系统的有效匹配。目前主流互联网公司凭借其网络开发优势,已开始在

表1　传统以太网与RDMA无损网络特点对比

对比类别	传统以太网	RDMA无损网络	
		IB	RoCE
产业链	开放度高，参与企业多	仅一家厂商	开放度高，参与企业多
产品规格	交换机：10Gb、25Gb、40Gb、100Gb 网卡：10Gb、25Gb	交换机/网卡：FDR（56Gb）、EDR（100Gb）、HDR（200Gb）	交换机/网卡：25Gb、100Gb
效率	多次数据内存拷贝，增加延迟；中断处理响应需CPU处理，占用CPU，影响性能	原生支持RDMA技术，支持内核旁路，性能高	支持RDMA技术，性能较高
网络配置	需要一定的网络配置保证网络互通	子网管理器自动配置，较为简便	在以太网基础上需要进行流控参数的配置
网络规模	组网灵活、适用各种规模和距离的组网需要	多为小规模局部专用网络	多为小规模局部专用网络
使用场景	应用场景广泛，通用性、兼容性、扩展性好	适合特定应用场景下需要高速互联网络的小规模组网，需要应用尽量支持RDMA	适合特定应用场景下需要高速互联网络的小规模组网，实现高性能需求应用尽量支持RDMA
特点	性能较低，适用性广	高带宽、低时延、传输操作低CPU消耗，难以兼容以太网	高带宽、低时延、传输操作低CPU消耗，兼容传统以太网
成本	广泛部署，成本低	需要端到端采购部署，供应商单一，成本较高	交换机和网卡需支持相关功能，供应商较多，成本较低

数据中心组网中采用白盒部署RDMA技术。综合考虑现网设备的成熟度和建设成本，近阶段无损网络的优势应用场景可分为高性能计算场景和高性能存储（全闪）场景。

2.1 高性能计算场景

针对现网运营中的近百个高性能计算场景组网的调研结果显示，约29%的组网采用IB方式以获取高处理性能。但IB组网的专网设备建设成本比以太网高出50%，且其集中控制的方法导致组网规模受限，特别是在国产化演进的要求下，IB组网并未展现出相较于RoCE组网更为显著的优势。为测试高性能计算场景下IB与以太网组网方式的性能情况，本文构建一个由全100GE和8台服务器构成的对等测试环境，进行超算测试。高性能计算场景组网测试示意如图2所示。

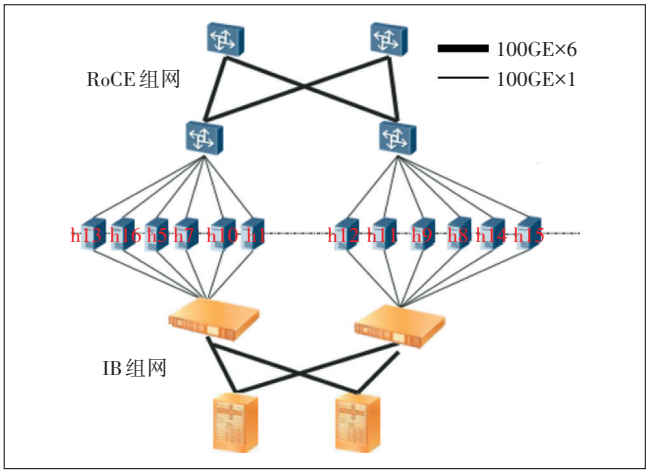

图2　高性能计算场景组网测试示意

在未开启RoCE技术的情况下，以太交换机时延为500 ns，IB交换机时延为90 ns，两者差距较大；在开启RoCE进行业务性能优化后，2种组网方式的性能差异可控制在10%以内，如表2所示。

表2　IB与以太网无损网络性能对比

高性能计算应用	相同任务完成时间/s		RoCE较IB性能提升相对比例
	RoCE	IB	
气象环境	5 282	5 087	3.83%
	42 832	42 866	持平
	2 000	1 888	5.93%
量子力学	315	314	0.32%
	121	121	持平
新材料模拟	5 418	5 393	0.46%
分子动力学	88.55	89.543	1.11%
Linpack	20 852.5	20 882.2	0.14%

2.2 高性能存储场景（全闪）

目前存储网络的组网方式主要包括FC、TCP以及RoCE等[8]。其中，虽然TCP方式有助于实现数据中心网络全IP化，但是如表1中所示，它存在时延高、CPU消耗高等硬性缺陷，不适用于高性能存储场景。为了对FC和RoCE面向高性能存储组网做比较测试，笔者模拟数据库业务，测试单并发的极致时延和单端口多并发的单业务最大性能，如图3所示。测试结果显示，RoCE可将每秒输入/输出操作（Input/Output Operations Per Second，IOPS）提升50%至100%，时延下降30%至50%。

RoCE全闪存储产业目前已较为成熟，在应用、操作系统、网卡、交换机、NVMe存储等全系列已有较多

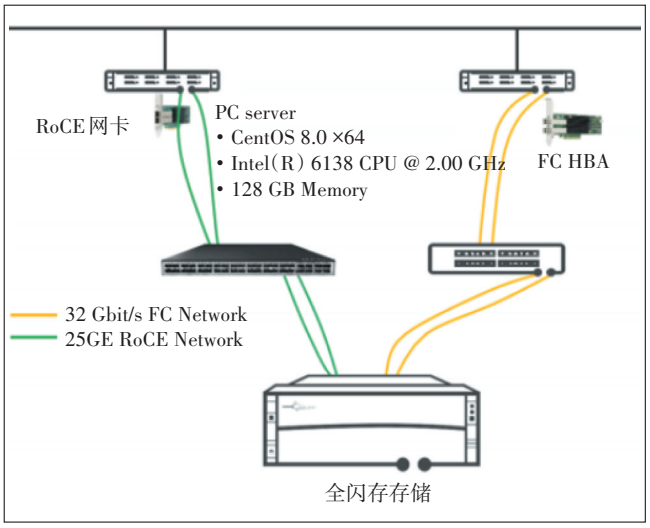

图3 高性能存储场景组网测试示意

的支持厂家,整个社区活跃度较高。从表3可以看出,RoCE在存储性能、带宽、管理方面有显著优势,但要替换FC连接全闪存,RoCE仍需从0丢包机制、秒级主备切换和存储即插即用特性等方面进一步提升。因此,综合考虑技术可拓展性、成本、性能、自主可控等多方面因素,建议高性能存储网络(全闪)逐步从FC、TCP向RoCE演进。

2.3 组网策略建议

建议将无损网络组网架构分为业务/计算和存储2个层面。对于业务/计算层面,建议延续现网常见的SDN组网架构;对于存储网络层面,建议采用无损交换机组网。

表3 FC与RoCE存储网络特性对比

类别	子类别	FC	RoCE
性能	带宽升级	32/64G	400G
	时延/μs	80	50
	0.1%丢包IOPS下降30%	稳定0丢包	拥塞易丢包
可靠性	升级中断时间	<1 s	<1 s
	存储故障,主备时间 金融应用要求<1 s	<1 s	<8~15 s
易用性	日常运维	智能定位	智能运维
	存储部署、一键开通	即插即用	手动配置
开放管理	统一管理	封闭架构、专人专管	开放以太、统一运维
	总体成本	TCO高	TCO低

针对小规模组网需求,无损网络的组网架构如图4所示。计算网络用于计算节点之间,以及计算和存储节点之间的通信,要求网络具备低时延和高吞吐能力,建议部署RoCE无损网络技术。业务管理网络用于集群管理、资源监控、作业管理等相关业务操作,建议部署传统TCP/IP网络。存储后端网络用于用户存储节点之间的后台通信,要求网络具备低时延、高吞吐能力,建议部署RoCE无损网络技术,RoCE交换机单归接入,不部署Leaf堆叠或MC-LAG,采用双平面组网。对于同城互通网络DCI-Leaf,建议采用8个100G

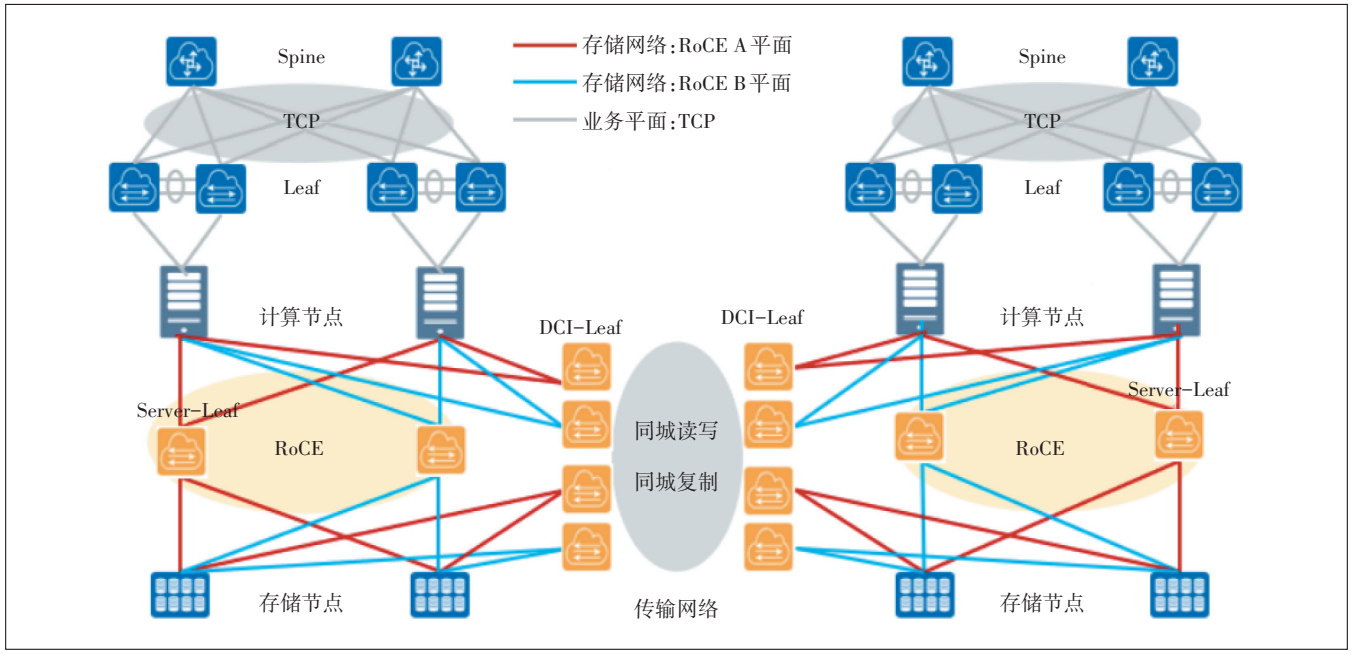

图4 小规模无损网络组网架构示意

上行以太网端口，DCI采用4×10/25G与波分设备进行互联，并根据同城距离设置不同等级的端口数量。

从小规模组网的实测情况来看，在开启PFC、ENC动态优化配置后，网络的时延、带宽性能明显提升，CPU消耗也有降低，这一结果验证了该组网方式可以满足高性能场景和网络演进的需求。

针对中大规模组网需求，无损网络的组网架构如图5所示。针对存储后端网络，建议存储节点可以直接挂在Spine上。当节点规模超过6 000个，或者要求存储具有更好的独立扩展性时，建议存储侧RDMA组网选用三级组网方式。

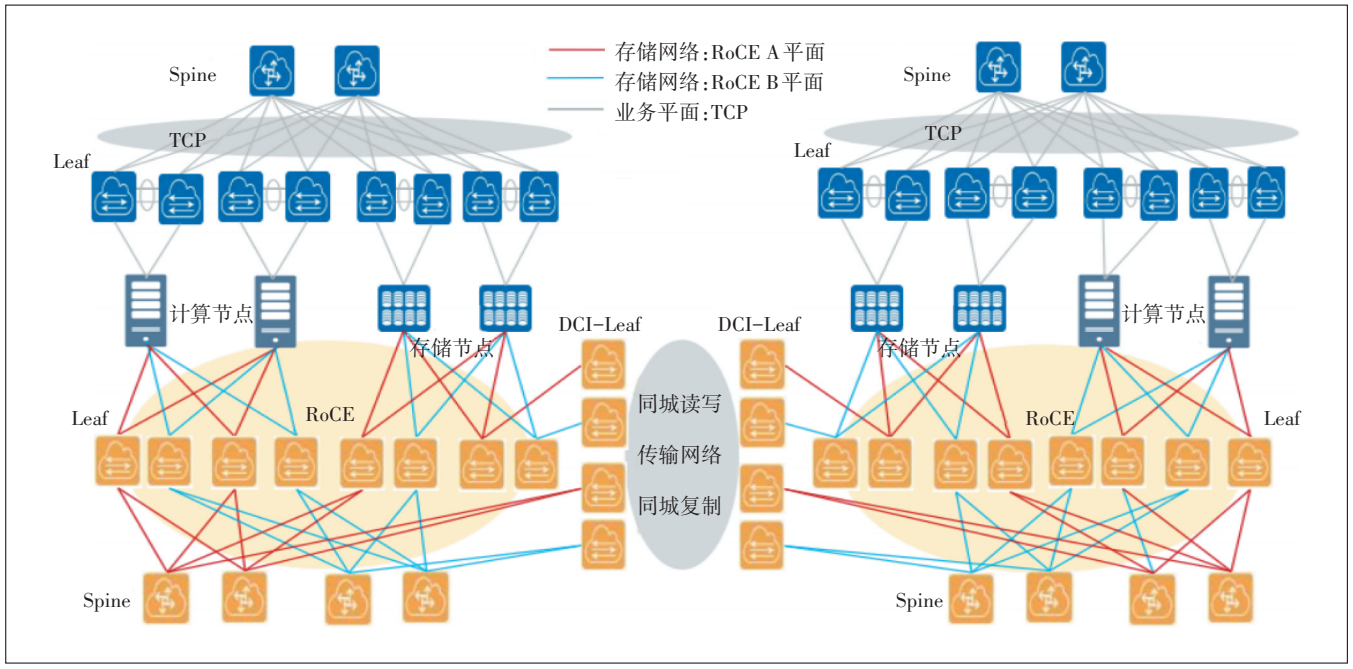

图5　中大规模无损网络组网架构示意

3 结语

无损网络是数据中心网络演进的新方向，是未来超融合以太网数据中心通用网络的关键技术。相较于传统以太网、FC等技术，RDMA技术由于其性能和较好的技术成熟度，在高性能计算和高性能（全闪）存储组网方面有较好的表现。无损网络组网可优先从上述2个方面切入试点，以积累无损交换机的运维经验。同时，总结本文测试经验，发现RoCE交换机的使用需要较高的技术开发掌控力，SDN网络的资源布局和开发能力也需要优化提升。

参考文献：

[1] 中国信息通信研究院. 中国算力发展指数白皮书[R/OL]. [2024-02-02]. http://www.caict.ac.cn/kxyj/qwfb/bps/202211/P020221105727522653499.pdf.

[2] 王少鹏，郑常奎，芦帅，等. 数据中心无损网络关键技术研究[J]. 信息通信技术与政策，2021(10)：68-74.

[3] 刘志锋，叶志伟，蔡敦波，等. RDMA技术研究综述[J]. 软件导刊，2022，21(12)：266-271.

[4] 吴莹. 一种网络设备拥塞控制机制的测试方法，系统及设备：CN114866477A[P]. 2022.

[5] 赵精华，郭亮. 智能无损网络：数据中心网络性能优化策略[J]. 中国电信业，2021(S1)：67-72.

[6] 华为技术有限公司. 智能无损网络技术白皮书[R/OL]. [2024-02-02]. https://www.163.com/dy/article/GIFT1JH90511BHI0.html.

[7] 范旭光. 无损网络数据中心应用概述[J]. 通信世界，2019(33)：36.

[8] 刘军，韩骥，魏航，等. 数据中心RoCE和无损网络技术[J]. 中国电信业，2020(7)：76-80.

作者简介：

蔡毅，毕业于同济大学，高级工程师，学士，主要研究方向为算力网络关键技术及部署实践、IT支撑网、云资源池、AIops等创新应用实践；樊蓉，毕业于中国科学院大学，高级工程师，硕士，主要研究方向为算力网络关键技术及部署实践、智慧城市相关信息基础设施关键技术及创新应用实践等；金沙，毕业于上海交通大学，工程师，学士，主要研究方向为IT支撑网、IT云资源池、分布式数据库技术等。