综述

面向国产算力的超大规模智算集群网络：关键挑战、技术途径与发展趋势

张慧峰[1]，刘宁春[1]，龙卫平[1]，陆平静[2]，邹涛[1]，隆克平[3]，张汝云[1]，朱俊[1]
（1. 之江实验室，浙江 杭州 311121；2. 国防科技大学计算机学院，湖南 长沙 410073；
3. 北京科技大学计算机与通信工程学院，北京 100083）

摘　要：随着大模型等人工智能技术的快速发展，构建超大规模智算集群网络成为必然需求。然而，我国在建设此类基础设施过程中，面临NVIDIA GPU短缺、算力资源成本高昂及利用率偏低三大核心困境。鉴于国产算力相较于NVIDIA的产品与技术生态尚未成熟，系统性分析了引入国产算力后智算集群网络将面临的三大关键技术挑战：国产智算集群网络互联能力的提升；智算集群网络传输效率的提高；智算集群网络可用性的增强。针对上述挑战，从网络架构、网络设备、通信协议到网络故障等方面，深入研究现有技术路径与解决方案，并结合实际集群建设经验，提出面向自主可控、高效可靠智算集群网络基础设施的未来发展趋势，为国产化大规模智算集群建设提供理论支撑与实践参考。

关键词：智算集群网络；国产算力；网络架构；网络设备；传输协议；集合通信
中图分类号：TP393；TN91
文献标志码：A
doi: 10.11959/j.issn.1000−0801.2025229

Hyperscale intelligent computing cluster networks for domestic computing power: critical challenges, technical pathways, and future trends

ZHANG Huifeng[1], LIU Ningchun[1], LONG Weiping[1], LU Pingjing[2],
ZOU Tao[1], LONG Keping[3], ZHANG Ruyun[1], ZHU Jun[1]
1. Zhejiang Lab, Hangzhou 311121, China
2. College of Computer Science and Technology, National University of Defense Technology, Changsha 410073, China
3. School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing 100083, China

Abstract: With the rapid advancement of artificial intelligence technologies such as large-scale models, constructing

收稿日期：2025−06−04；**修回日期：**2025−09−22
通信作者：朱俊，zhu_j@aliyun.com
基金项目：国家自然科学基金资助项目（No.U22A2005）；国家重点研发计划项目（No.2022YFB2901500，No.2022YFB2901400，No.2023YFB2903900）；浙江省重点研发计划项目（No.2024SSYS0001）
Foundation Items: The National Natural Science Foundation of China (No.U22A2005), The National Key Research and Development Program of China (No.2022YFB2901500, No.2022YFB2901400, No.2023YFB2903900), Key Research & Development Program of Zhejiang (No.2024SSYS0001)

ultra-large-scale intelligent computing clusters has become an imperative. However, China faces three core challenges in building such infrastructure: shortages of NVIDIA GPUs, prohibitively high costs of computing resources, and their chronic underutilization. Given the immaturity of domestic computing solutions relative to NVIDIA's established product and technological ecosystem, three critical technical challenges for intelligent computing cluster networks upon adopting domestic alternatives were systematically analyzed: enhancing interconnectivity capabilities within domestic computing clusters; improving data transmission efficiency across intelligent computing networks; and strengthening network availability guarantees. To address these challenges, an in-depth examination of existing technical approaches and solutions was conducted, spanning from network architecture and network devices to communication protocols and network fault tolerance. Drawing on practical cluster deployment experience, the future development trajectories toward building an autonomous, controllable, efficient, and reliable intelligent computing network infrastructure were further outlined. The theoretical foundations and practical references for large-scale domestic computing cluster construction were provided.

Key words: intelligent computing cluster network, domestic computing power, network architecture, network device, transmission protocol, collective communication

0 引言

随着人工智能（artificial intelligence，AI）大模型技术的迅猛发展，其在各行各业中的广泛应用正深刻地改变着生产力的发展模式和产业格局。为满足这一趋势下对算力资源的巨大需求，构建一个高可靠、高效、高性价比、绿色节能的超大规模智算集群基础设施已成为当务之急。近年来，全球顶尖科技公司和云服务提供商在构建超大规模智算集群方面取得了显著进展，以满足日益复杂的AI模型训练需求。

NVIDIA在智算集群网络方面占据主导地位，通过其高性能图形处理器（graphics processing unit，GPU）（如A100、H100等）与NVLink、无限带宽（InfiniBand，IB）网络等协议构建大规模AI算力集群。NVIDIA的Spectrum-X平台[1]被多个云服务提供商和企业采用，用来提高AI基础设施的网络性能。2020年，微软与OpenAI合作构建了Azure AI超算平台[2]，该平台拥有超过1万个NVIDIA A100 GPU，每个GPU拥有400 Gbit/s网络带宽且由IB网络相互连接。Amazon建设了Amazon EC2 UltraClusters AI集群网络[3-4]，该网络包含4 000个NVIDIA A100 GPU的EC2 P4d集群、2万个NVIDIA H100 GPU的EC2 P5集群、3万个Trainium芯片（Amazon专为训练AI模型而设计的芯片）构成的EC2 Trn1n集群等。特斯拉的xAI项目超级计算机集群“Colossus”已达到10万个NVIDIA H100 GPU的规模，采用基于NVIDIA Spectrum-X以太网的远程直接内存存取（remote direct memory access，RDMA）网络互联架构。Meta[5-6]已建成基于融合以太网的远程直接内存访问（RDMA over converged Ethernet，RoCE）的2.4万个NVIDIA H100 GPU卡AI集群。Google Cloud AI HypercomputerAI[7]训练集群可部署10万个Trillium芯片。Trillium芯片[8]是Google Cloud推出的第六代张量处理单元，专门被设计用于处理AI工作负载。

当前，国内智算集群网络建设正处于快速发展阶段，各大科技公司和研究机构正积极部署高性能的网络基础设施以支持大规模AI模型训练和数据处理需求。

DeepSeek[12]通过互联两个两层Fat-Tree网络，采用IB网络协议和Mellanox QM8700 IB交换机（40×200 Gbit/s）部署了拥有1万个高速串行计算机扩展总线标准（peripheral component interconnect express，PCIe）A100 GPU的Fire-

Flyer2 集群。阿里云 HPN 7.0[13]采用了双平面架构、轨道优化技术，能够用二层网络架构在单个 Pod 内互联 1.5 万个 GPU，用 3 层网络架构互联 10 万个 GPU。华为昇腾 910B 集群采用单轨接入、二层 Spine-Leaf 架构，可互联 9 216 个、18 432 个甚至 147 456 个昇腾 910B 卡。字节跳动 MegaScale[15]搭建了超过 1 万个 NVIDIA Ampere GPU 的 AI 训练集群。百度智能云采用 IB 交换机搭建了 IB 集群网络，可支撑 1.6 万卡的规模，配置 NVIDIA A100 GPU 卡[17]。目前，百度百舸平台已在建设异构混合 GPU 集群，包含 NVIDIA GPU、昆仑芯、昇腾等算力卡，目标是达到 10 万异构卡集群的规模。腾讯 AI 集群网络已从星脉网络 v1.0 升级到 v2.0，其采用的网络交换机从 25.6 Tbit/s 升级到 51.2 Tbit/s，光模块从 200 GB 升级到 400 GB。星脉网络 2.0 可支持单集群 10 万卡 GPU 以上的规模。零一万物沿用 DGX SuperPod 的网络配置，采用 IB 三层组网架构，理论支持 6.4 万 GPU 集群（当前规模千卡级别）。无问芯穹 Infini-AI 云平台已集成大模型异构千卡混训能力，具备万卡扩展性，支持包括 AMD、NVIDIA 以及国产华为昇腾、天数智芯、沐曦、摩尔线程 6 种异构芯片在内的大模型混合训练。此外，国内各大厂商如壁仞科技，运营商中国移动[16]、中国电信、中国联通，相关科研机构等也在相继建设 AI 智算集群网络，旨在构建具有我国自主知识产权的异构智算集群网络。

当前面向国产算力的超大规模智算集群网络才初步显现。为了满足日益增长的大模型训练任务需求，集群网络必将持续面临一系列的技术挑战。

1　现状与挑战

面向国产算力的超大规模智算集群网络面临以下三大困境。

（1）高性能算力芯片短缺。由于美国对华实施的半导体出口管制措施（美国商务部工业和安全局对《出口管制条例》进行修订，先后形成“1007 规则”和“1017 规则”，将高性能芯片及半导体制造设备相关物项加入商业管制清单[9]），国内智算集群建设所依赖的高端 GPU 芯片存在“卡脖子”风险。

（2）算力使用成本高昂。GPU 服务器及相关算力资源的价格高昂，使得构建大规模智算集群所需要的资金投入巨大，从而阻碍了低成本算力服务的普及，如一台 NVIDIA H100 显卡模组现货价格在 200 多万元。

（3）算力资源利用率低下。当前智算集群的算力资源利用率普遍偏低，特别是模型计算利用率（model flops utilization，MFU）等关键指标，如字节跳动的 MegaScale 系统，在 12 288 个 GPU 上训练 175 亿个参数的语言模型时实现了 55.2% 的 MFU[15]，DeepSeek[10,12,14]也同样为提高 MFU 提出一系列软硬件协同优化技术[12]。目前，国内涌现出众多的国产算力产品，如昆仑芯、沐曦、寒武纪、壁仞、摩尔线程、燧原、国科微、无问芯穹、天数智芯等。尽管如此，相较于 NVIDIA 成熟的产品线及其 CUDA 生态系统的完备性，当前各类国产算力解决方案尚处于发展阶段，存在一定的技术和市场成熟度差距。这为面向国产算力的大规模智算集群带来了新的挑战，特别是在网络互联方面，如何实现异构国产算力之间的高速、高效、稳定通信成为一个亟待解决的问题。

因此，建设超大规模国产算力集群网络主要面临以下三大挑战。

1.1　挑战 1：提升国产异构智算集群网络互联能力

提升超大规模国产异构算力智算集群网络互联能力的相关技术如下。

（1）网络设备性能提升。网络设备的能力（带宽、端口数、端口速率等）很大程度上决定了能支持 GPU 卡的规模；另外，网络设备须支持异构算力（如 NVIDIA GPU、华为昇腾、天数智

芯、沐曦、摩尔线程等）互联。

（2）网络拓扑结构优化。网络拓扑决定了节点间的连接方式，影响数据传输路径的选择、时延及整体网络性能。如Fat-Tree拓扑可以在一定程度上减少通信瓶颈，提高网络吞吐量，而环形或网格状结构则有助于降低平均跳数，缩短通信时延。同时，随着集群规模的增长，网络拓扑应具备扩展能力，支持灵活调整以适应不同的应用和服务需求。

（3）集合通信的兼容性。不同国产厂商的算力采用各自的集合通信库，如华为集合通信库（Huawei collective communications library，HCCL）、阿里云集合通信库（Alibaba collective communications library，ACCL）、腾讯集合通信库（Tencent collective communication library，TCCL）、百度集合通信库（Baidu collective communication library，BCCL）等，需要实现集合通信在异构算力集群中通用性，满足数据交换要求。因此，提高网络互联能力须综合考虑网络设备的选型与配置、合理的网络拓扑结构以及统一集合通信模式等多个方面的工作。

1.2 挑战2：提高智算集群网络传输效率

提高超大规模异构国产算力智算集群网络的通信效率，面临下述问题。

（1）网络负载不均。由于大模型训练流量具有低熵的特性[6]，容易导致哈希极化，造成某些热点网络节点成为瓶颈，极大限度降低了整个系统的效率。

（2）网络拥塞问题。当大量数据同时通过有限带宽的链路时会发生网络拥塞现象。

（3）集合通信不畅。大模型训练任务依赖于集体通信操作（如All-Reduce、All-Gather等），但它们也容易成为通信性能瓶颈，大大增加训练整体执行时间。

（4）在网计算能力缺失。在网计算指在网络基础设施内部直接进行部分计算处理。然而，当前在网计算技术主要由国际巨头主导，国产厂商在这一领域起步较晚，对于大规模异构智算集群而言，缺乏足够的在网计算支持意味着更多的计算任务必须回退到端点设备上完成，将增加报文时延和通信瓶颈。因此，提高智算集群网络传输效率不仅要解决传统意义上的网络传输效率问题，还需要引入新型网络传输技术和协议、开发智能化的管理和调度系统以及提高网络自身的计算和服务能力。

1.3 挑战3：增强智算集群网络可用性

增强超大规模异构国产算力智算集群网络运行的可用性和可靠性，是构建稳定、高效计算环境的关键挑战，目前面临着以下问题。

（1）硬件设备故障。温度变化等因素可能导致光模块失效，进而影响通信质量；智能网卡或RDMA网卡出现故障，会导致节点间的通信中断，影响任务进度；交换机承载着大模型训练流量转发任务，配置错误或硬件故障都可能引发广泛的网络瘫痪。

（2）传输协议复杂性和互操作。网络传输协议（如RoCE等）复杂度高、配置难度大，不正确的配置可能导致性能下降甚至连接失败。不同厂商提供的网络设备和服务之间可能存在兼容性问题，特别是在混合使用多种技术和标准的情况下。

（3）网络架构设计不合理。如集群网络拓扑结构设计未能充分考虑负载均衡、扩展性和容错能力，可能导致局部热点形成或整体性能瓶颈。

（4）故障检测定位难和恢复慢。缺乏可视化工具和管理平台，运维人员很难迅速识别并定位故障源。另外，故障具有复杂的依赖关系，当一个组件出现故障时，其影响范围可能波及多个其他部分。传统的故障处理方式往往依赖人工判断和操作，不仅耗时且容易出错。

针对上述3个技术挑战，本文研究了多个典型智算集群网络，结合国产算力资源的现状和特点，提出了总体应对思路和方法，从网络拓扑、

硬件设备、传输协议、集合通信、负载均衡、拥塞控制、在网计算、状态监测、故障检测、故障恢复等技术方向入手，给出当前主要技术途径，并进一步讨论未来技术发展的趋势。

2　总体思路

基于国产算力的超大规模智算集群网络可以划分为网络架构、网络设备、通信协议与传输优化、网络故障检测与恢复 4 个关键技术领域，这些技术方向对集群网络性能、可靠性和效率至关重要，总体应对技术思路如图 1 所示。

提升智算集群网络互联能力主要从网络拓扑结构、网络互联方式、集合通信库、高性能交换机、智能网卡、光模块等方面探索解决途径。选择合适的网络拓扑结构（如 Fat-Tree、Dragonfly 等）、网络互联方式，如基于中央处理器（central processing unit，CPU）、GPU 的互联，能有效减少通信时延并提高带宽利用率。异构计算卡之间的通信还需要设计统一的集合通信库。网络交换机通过提供高带宽、低时延的端口，支持大规模数据传输，对交换机的选型也直接决定网络的规模与通信能力。另外，网络交换机还需要提供先进的流量管理功能，如优先级队列和拥塞控制，确保关键任务数据优先传输。智能网卡通过卸载主机的网络处理任务，减少数据传输中的 CPU 干预，提升网络通信效率。光模块通过提供高带宽和长距离传输能力，支持大规模数据传输，确保低时延和高可靠性。此外，低功耗和高密度的光模块有助于优化数据中心能效。

提高智算集群网络传输效率主要从拓扑结构、传输协议、负载均衡、拥塞控制、集合通信库、在网计算等方面探索解决途径。优化拓扑结构（如轨道优化等）能够减少数据传输跳数，降低时延并提高带宽利用率。合适的传输协议，如传输控制协议（transmission control protocol，TCP）、用户数据报协议（user datagram protocol，UDP）、RoCE、IB，可以显著提升传输效率，如 RoCE 和 IB 支持 RDMA 技术，绕过操作系统内核直接访问内存，大幅降低时延并提高吞吐量。优化协议参数（如窗口大小、超时重传机制）也能

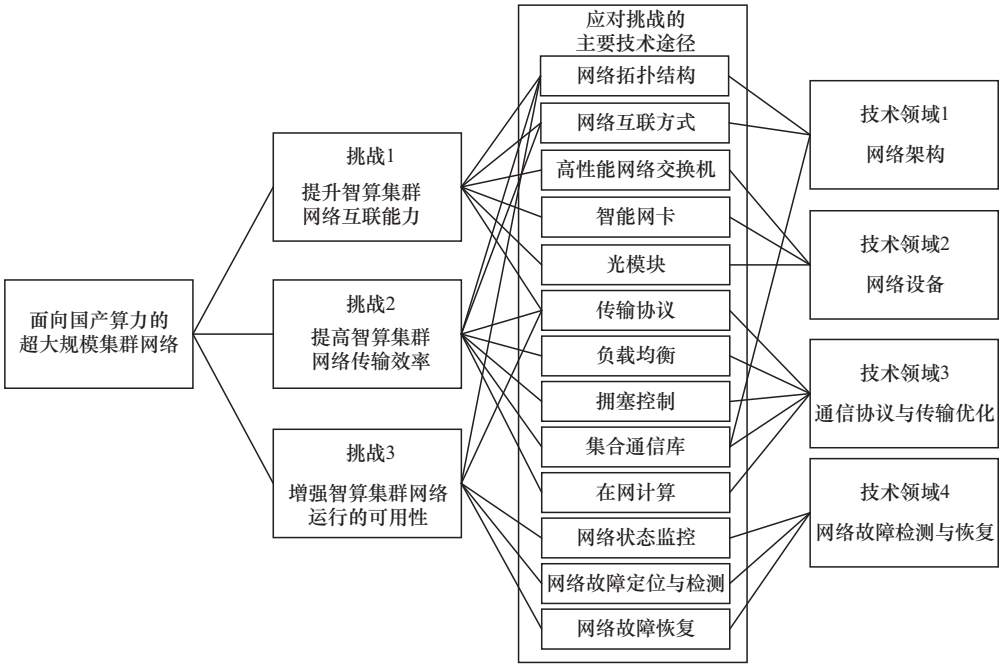


图 1　总体应对技术思路

进一步提升性能。负载均衡通过动态分配网络流量，避免某些链路或节点过载，从而提高整体传输效率。拥塞控制机制，如显式拥塞通知（explicit congestion notification，ECN）、数据中心量化拥塞通知（data center quantized congestion notification，DCQCN）通过检测和缓解网络拥塞，避免数据包丢失和重传。优化集合通信库（如HCCL、ACCL、BCCL等）能够提升多节点协同计算的效率，减少通信开销并提高并行计算性能。在网计算通过在网络设备中执行部分计算任务（如数据聚合、压缩），减少数据传输量和主机处理负担，从而降低时延并提高整体效率。

增强智算集群网络运行的可用性主要从网络状态监控、网络故障检测、网络故障恢复、拓扑结构设计等方面探索解决途径。智能的网络状态监控，包括但不限于流量分析、性能指标跟踪、硬件健康检查等，可帮助运维人员及时了解网络运行状况。通过部署分布式监控代理，收集来自各节点的数据，并将信息汇总至管理平台进行统一处理。建立快速而准确的网络故障检测机制是保障网络可靠性的重要环节，可结合硬件级别的网络错误报告和软件层面的日志记录来识别异常情况。智能诊断工具自动化地分析日志文件和其他相关信息，定位故障根源并提供修复建议。完善的网络故障恢复策略确保在网络遇到问题后能够迅速恢复正常运作，引入网络故障恢复自愈相关技术，如自修复交换机端口和自动化的配置回滚机制缩短故障响应时间，增强网络的整体弹性。另外，设计高冗余的拓扑结构，确保即使某些节点或链路出现故障，数据依然可以通过代替路径传输，维持网络连通性。

3 当前技术途径

围绕总体应对技术思路，本文对每一类技术领域当前主要技术途径展开论述。

3.1 网络架构

3.1.1 网络拓扑结构

数据中心典型拓扑结构如图2所示，智算集群网络主要的拓扑结构有：Clos、Fat-Tree、Spine-Leaf、Dragonfly、Slim Fly、Torus。Clos拓扑结构[18]是一种多层网络交换机的无阻塞连接方式，通过将交换机分层，每一层都与上一层和下一层的交换机相连，实现了高度的可扩展性和灵活性。Fat-Tree和Spine-Leaf拓扑结构是典型的Clos网络的代表。Fat-Tree[19]（如图2（a）所示）由核心层、汇聚层和接入层3个层次组成，在传

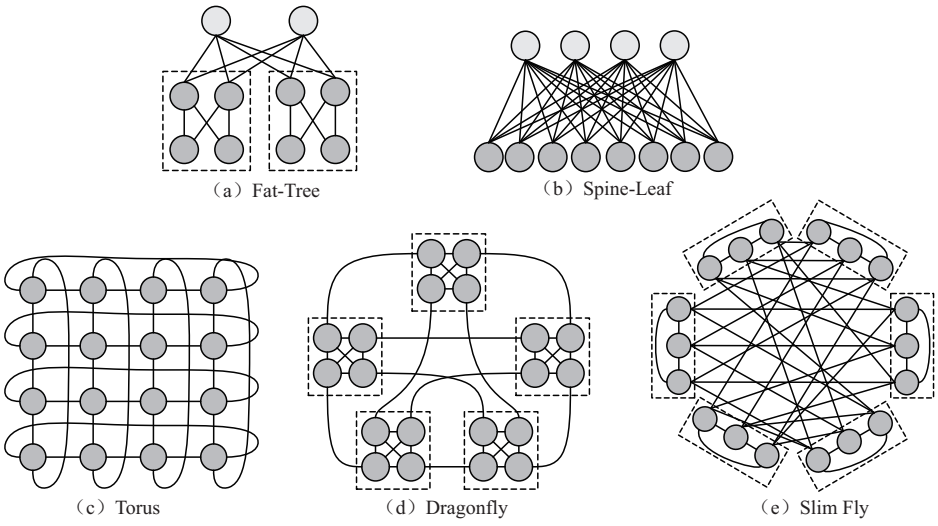


图2 数据中心典型拓扑结构

统树状结构的基础上增加冗余副本，使每一层都具有相同的聚合带宽，从而可利用低成本的设备构建大规模的无阻塞网络。Spine-Leaf[21]（如图 2（b）所示）基于扁平化设计，由 Spine 层和 Leaf 层组成，每个 Leaf 交换机直接连接到所有 Spine 交换机，形成一个全互联的网络结构。Leaf 交换机负责连接物理服务器，而 Spine 交换机负责互联所有的 Leaf 交换机，使任意两个服务器之间的连接都经过相同数量的设备，保证时延是可预测的。Torus[25]（如图 2（c）所示）作为高性能计算中被广泛应用的互联架构，采用 n 维环面结构实现节点周期性连接，每个节点通过双向通道与相邻 $\pm1(\mathrm{mod}\ k)$ 节点直连，兼具终端与路由功能。其核心优势是物理拓扑与逻辑路径的高度一致性保障了通信局部性最优，良好的路径多样性实现了负载均衡，双向通道配置较单向网络可降低平均跳数。通过维度参数 n 的调节，网络吞吐量可线性增长至对分带宽极限，形成性能与复杂度的可控平衡。

Dragonfly[22]（如图 2（d）所示）是一种高性能网络拓扑结构，以小网络直径和低成本为技术特点，适用于高性能计算场景。该拓扑结构分为：Switch 层、Group 层和 System 层，每一层都通过特定的参数进行描述和连接。这种结构通过优化路由和连接方式，显著降低大规模组网中的跳数，从而提升网络性能。它被运用在 Cray XC 系列超级计算机中，如 Piz Daint、Trinity 和 Cori 等。Slim Fly[23]（如图 2（e）所示）是一种高性能且成本效益高的网络拓扑结构，具有接近理论上最优的网络直径。相比于其他网络拓扑（如 Fat-Tree、Clos 等），Slim Fly 通过优化网络结构以降低网络直径，减少传输时延、成本和功耗[24]。该结构允许构建具有超过 10 万个节点的全带宽网络，直径仅为 2。

网络拓扑设计的核心在于平衡性能、成本与可扩展性的矛盾。不同拓扑结构的诞生均针对特定场景的通信瓶颈。Fat-Tree[19-20]基于冗余设计理念，为解决传统树状拓扑的带宽阻塞问题，通过分层冗余（核心–汇聚–接入层全互联）实现无阻塞通信，其本质是牺牲布线成本换取高吞吐量（聚合带宽=接入带宽），适用于高并发场景。Spine-Leaf[21]的扁平化革新，突破多层架构的时延不可控缺陷，采用两级全互联（Leaf 直连 Spine），其设计理念源于“缩短路径跳数”，确保任意服务器间通信仅需 2 跳（Leaf→Spine→Leaf），实现时延可预测性。Dragonfly/Slim Fly[22-24]对直径进行优化，面向超算场景的通信局部性需求，通过分组直连（Group 内全互联+Group 间部分互联）将网络直径压缩至 2~3 跳，其设计核心是“以拓扑换性能”，降低长距离通信开销。Torus[25]的维度扩展性，基于 n 维环面周期性连接，最大化利用物理邻近性，其设计理念强调“局部优先路由”，通过维度参数 n 线性提升对分带宽，适用于计算密集型负载。

在网络拓扑结构的基础上进一步进行组网优化，目前有以下两种方法。

（1）单轨与轨道优化

网络拓扑结构优化如图 3 所示。单轨网络结构是指在集群网络中每个节点或服务器通过单个网卡连接到网络，如图 3（a）所示。这种结构简单，成本较低，但在网络带宽或容错性方面存在局限性。单轨网络意味着每个节点只连接一个交换机，在网络出现故障时可能影响整体的可靠性和性能。Meta[5]基于 RoCE 的 AI 集群采用单轨技术通过三层 Clos 网络连接；华为昇腾 910B 集群网络架构采用单轨技术通过二层 Spine-Leaf 连接；DeepSeek[12]Fire-Flyer2 集群网络采用单轨方式部署两个两层 Fat-Tree 的网络拓扑。

轨道优化网络结构是指在集群网络中，每个节点或服务器通过多个网卡连接到网络，可提供更高的带宽和更好的容错性，如图 3（b）所示。轨道优化设计允许网络流量在多个路径上分布，

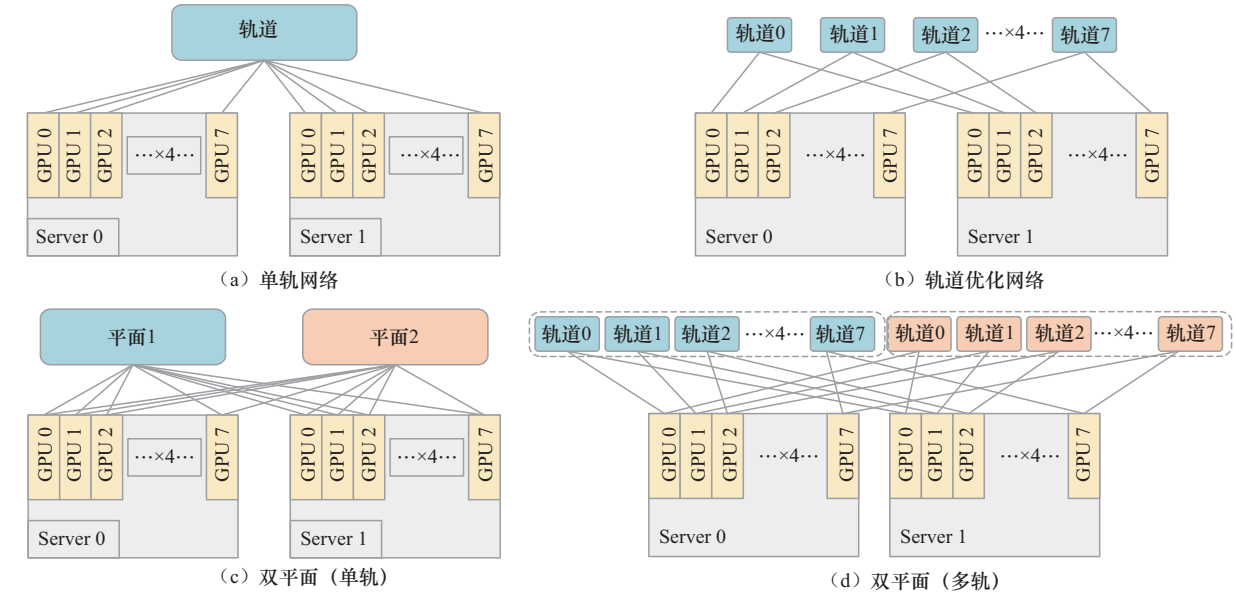


图3　网络拓扑结构优化

从而提高网络的冗余度和可靠性。如每个服务器中的不同GPU可以连接到不同的交换机，不同服务器的同号卡连接到相同的交换机，这样既提升了同号卡间的通信效率，也确保了网络的高可用性。Meta[11]采用Rail-only的方式，去除了传统GPU集群中的Spine层，只在同一个轨道内的GPU之间提供网络连接；NVIDIA SuperPOD[1]基于IB的AI集群采用轨道优化的技术通过3层Clos网络连接；字节跳动[15]、百度智能云[17]、腾讯云均采用轨道优化的技术通过3层的Fat-Tree网络拓扑搭建AI集群。

（2）单平面与双平面

单平面网络架构是指在一个网络层面上实现所有数据传输和通信，这种设计简化了网络结构，但可能在面对大规模数据传输和高负载时存在性能瓶颈。

双平面是一种创新的网络架构，通过构建两个完全相同的网络平面来优化大规模GPU集群的互联。每个GPU都对应两个网卡端口，两个端口分别连接不同的网络平面，如图3（c）所示。这两个网络平面的拓扑结构完全相同，且没有任何交叉连接。这种设计大幅降低了哈希极化的概率，优化了通信效率。双平面架构不仅提升了网络性能，还增强了网络的稳定性和可靠性，即使单一设备或单一平面出现网络故障，也不会影响整个集群网络的正常运行。连接双平面的链路称为双上联。阿里云[13]采用轨道优化和双平面的技术，通过3层Clos网络构建的AI集群可支持10万卡互联，如图3（d）所示。

单平面与双平面组网的性能指标见表1。目前，国内外的集群网络组网主要采用网络拓扑结构+单/多轨道+单/双平面组合的方法，具体采用哪一种方式组网须综合考虑训练任务性质、交换机性能、并行策略、可扩展性、容错性等多方面因素。

3.1.2　网络互联方式

网络互联技术包括机内互联与机间互联。

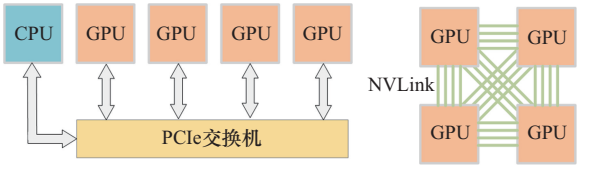
GPU互联技术如图4所示。当前的机内互联技术有PCIe[26]、NVLink[26]、UALink[27]、OLink等技术。PCIe（如图4（a）所示）是GPU与CPU通信的基础技术，具有通用性强和成本低的优势。目前已发布PCIe3.0（8 GT/s，单通道1 GB/s）、PCIe4.0（16 GT/s，单通道2 GB/s）、PCIe5.0（32 GT/s，单通道4 GB/s）、PCIe6.0

表1　单平面与双平面组网的性能指标

网络	容错能力	网络拥塞和性能	成本和管理复杂性	可扩展性
单平面	如果链路或交换机出现故障，可能会导致整个GPU或服务器失去网络连接	在高负载情况下，单平面可能成为通信瓶颈，限制数据传输速率，影响集群的整体性能	成本较低，因为需要的网络设备和链路较少，管理维护相对简单，网络结构更简单	在集群规模较小时可能表现良好，但随着集群规模的扩大，可能会遇到扩展性问题，因为所有通信都依赖于单一路径
双平面	双平面设计通过为每个GPU或服务器提供两个独立的网络路径	分散网络流量，减少单个链路的负载，提高通信效率，从而提升集群的整体性能	需要更多的网络设备和链路，导致初期硬件投入成本增加；网络的管理和维护也可能更复杂，因为需要监控和维护更多的网络组件	提供了更好的可扩展性，因为它允许网络在不影响性能的情况下扩展，同时也更容易适应集群规模的增长

（64 GT/s，单通道8 GB/s）。PCIe带宽有限，特别在连接多个GPU时，容易成为系统性能瓶颈，但随着PCIe技术的不断演进，PCIe7.0（128 GT/s）版本已于2025年发布，有望通过低成本的方式满足大规模GPU间的互联需求，为异构计算提供可持续升级的互联底座。NVLink（如图4（b）所示）是NVIDIA推出的高速互联技术，用于实现GPU芯片间低时延、高带宽的数据互联。从第一代到第五代，NVLink不断提升互联带宽和拓扑结构的灵活性。如单个NVIDIA H100 GPU支持多达18个NVLink连接，总带宽为900 GB/s，是PCIe5.0带宽的7倍。UALink是由AMD、博通、思科等联合推出的开放式互联标准，用于支撑可扩展的高性能GPU集群架构。该标准通过灵活的多层级互联设计，支持单集群内集成最多1 024个GPU加速器，形成跨机架协同的计算单元，突破NVIDIA封闭式方案（NVLink）的扩展限制，为大规模模型训练与推理提供开放、高效的硬件互联底座。英特尔牵头成立的CXL联盟推出了CXL2.0、CXL3.0规范，旨在通过高速、低时延的互联协议，实现AI高性能计算场景中CPU与GPU等的互联。中兴通讯提出的OLink开放交换互联协议，通过自研大容量交换芯片实现节点内GPU之间的高速交换互联，突破机内8个GPU限制，支持16~128个GPU超级计算节点的超大规模的GPU计算能力。

机间互联技术当前主要基于RDMA架构实现，其技术路线可细分为IB和基于以太网的RoCE。在超大规模智算集群场景中，IB网络凭借其无损传输特性和成熟的拥塞控制机制，可提供超低时延和高达800 Gbit/s的端到端连接，已广泛应用于高性能计算场景的机间互联。而RoCE技术通过将RDMA协议栈移植到以太网架构，在兼容传统数据中心网络基础设施的同时，通过引入流量控制和显式拥塞通知机制，有效降低了通信时延，为计算节点提供了更具成本效益的互联方案。此外，智算集群的机内和机间互联可能将统一采用开放的互联协议，如UALink，通过光电协同传输技术实现计算、存储、网络资源的全维度高速互联，从而构建支撑大模型训练的网络环境。



（a）基于CPU通过PCIe技术互连　（b）基于NVLink技术互连

图4　GPU互联技术

3.1.3　集合通信库——解决异构算力间的互通性问题

不同厂商的集合通信库通常针对自身硬件和网络架构进行了深度优化，导致通信库之间存在显著差异，造成异构算力间无法互通的问题。

（1）通信库各异，各厂商的通信库（如HCCL、ACCL、TCCL、BCCL等）在接口、协议和优化策略上各不相同，难以直接兼容。

（2）跨芯片互联困难：不同芯片（如NVIDIA

GPU、昇腾、昆仑芯、沐曦、摩尔线程等）间的通信优化难度大，缺乏统一的接口和协议。

智源研究院发布了开源的异构统一通信库FlagCX，旨在实现不同芯片之间的高效通信和大规模自适应通信优化。FlagCX提供统一的通信算子接口层，屏蔽底层不同实现细节，支持多种深度学习框架，通过标准化适配接口，复用芯片厂商原生通信库。百度BCCL基于NVIDIA集合通信库（NVIDIA collective communications library，NCCL）进行了功能扩展和优化，支持NVIDIA GPU、昆仑芯等标准计算卡的互联互通。BCCL通过CPU转发实现跨昇腾910B子集群和NVIDIA GPU子集群的连接，并通过Accelerator抽象屏蔽硬件差异，将芯片算子与上层策略解耦，确保不同芯片在百度百舸平台上达到高运行效率。无问芯穹发布的集合通信库IHCCM支持基于CPU或GPU的通信方式，解决了不同类型芯片之间的通信问题。

3.2 网络设备

相关网络设备主要包含高性能交换机、智能网卡、光模块等关键网络设备。

3.2.1 高性能网络交换机

数据中心高性能网络交换机见表2。为支持超大规模智算集群网络数据中心，国内外巨头纷纷研制出自研高性能交换机，目前主流的端口速率已经达到400 Gbit/s，华为的CloudEngine16800系列交换机最大支持576个400 Gbit/s端口、9 642 Tbit/s交换容量，CloudEngineXH9230系列液冷盒式交换机支持128个400 Gbit/s端口、51.2 Tbit/s交换容量。H3C S9827系列交换机支持128个400 Gbit/s端口。云尖信息SD8882-64D与SD8882-128D交换机分别提供64个400 Gbit/s端口和128个400 Gbit/s端口。锐捷网络RG-S6990-128QC2XS高密度盒式交换机支持128个400 Gbit/s端口。中兴通讯发布了ZXR10 9900X系列交换机提供400 Gbit/s端口，支持向800GE平滑演进。

表2 数据中心高性能网络交换机

交换机厂商	系列	交换机型号与吞吐量
NVIDIA	NVIDIA Quantum-X800 IB network	NVIDIA Quantum-X800 Q3400 144×800 Gbit/s
	NVIDIA Spectrum-X800 Ethernet platform	NVIDIA Spectrum SN5600 64×800 Gbit/s/128×400 Gbit/s
Cisco	Cisco Nexus 9000系列	Cisco Nexus 9808 288×400 Gbit/s576×200 Gbit/s
Juniper Networks	PTX系列	PTX10002-36QDD 36×800 Gbit/s
	QFX系列	QFX5240-64QD 64×800 Gbit/s
Broadcom	Tomahawk系列	Tomahawk®5 128×400 Gbit/s
华为	—	CloudEngine16800 576×400 Gbit/s CloudEngineXH9330 128×400 Gbit/s
云尖信息	SD8882系列	SD8882-64D 64×400 Gbit/s SD8882-128D 128×400 Gbit/s
锐捷	高密度盒式交换机	RG-S6990-128QC2XS 128×200 Gbit/s/128×400 Gbit/s QSFP112端口
	高密度框式交换机	RG-N18010-XH 288×400 Gbit/s/384×100 Gbit/s
		RG-N18018-XH 276×400 Gbit/s/768×100 Gbit/s
新华三	H3C S9827系列	H3C S9827-128DH 128×400 Gbit/s QSFP112端口
		H3C S9827-64EP 64×800 Gbit/s OSFP800端口
		H3C S9827-64E 64×800 Gbit/s QSFP-DD800端口
中兴	ZXR10 9900X系列交换机	ZXR10 9900X支持高密度400 Gbit/s端口，支持向800 Gbit/s平滑演进

同时国内已经推出了 800 Gbit/s 端口速率的交换机，如华为的 100 Tbit/s（128×800 Gbit/s）盒式以太交换机 CloudEngineXH9330、H3C 的 S9827 交换机（64 × 800 Gbit/s），可满足超大规模数据中心和 AIGC 算力网络高密度服务器接入的组网需求。

在国际上，NVIDIA Quantum-X800 IB Network 中 Q3400 交换机提供 144 个 800 Gbit/s 端口，Spectrum-X800 平台中的 Spectrum SN5600 交换机提供 64 个 800 Gbit/s 端口、51.2 Tbit/s 交换容量，并且支持单个 2U 交换机中最多 128 个 400 Gbit/s 以太网端口或 64 个 800 Gbit/s 接口。思科 Cisco Nexus 9000 系列交换机提供 400、800 Gbit/s 端口速率。Cisco Nexus 9808 交换机提供 288 个 400 Gbit/s 端口。瞻博网络 Juniper Networks PTX 系列采用 Express 5 芯片，PTX10002-36QDD 支持 36 个 800 Gbit/s 端口，QFX 系列提供高密度的 400 Gbit/s 和 800 Gbit/s 连接，QFX5240-64QD 具有 64 个 800 Gbit/s 端口。博通 Tomahawk 系列芯片支持的交换机端口从 100 Gbit/s 演进到 200 Gbit/s 再到 400 Gbit/s，整机最大支持 128 个 400 Gbit/s 端口。

随着智算需求的增长，网络设备的端口速率和交换容量将进一步提升，交换机端口速率从 200 Gbit/s、400 Gbit/s 向 800 Gbit/s、1.6 Tbit/s 提升，交换芯片带宽容量从 25.6 Tbit/s、51.2 Tbit/s 向 102.4 Tbit/s 跨越。目前，800 Gbit/s 交换机开始放量，102.4 Tbit/s 交换芯片已于 2025 年推出。AI 大模型参数量持续增长倒逼集群组网规模提升，叠加 AI 芯片带宽提升，促使交换机端口速率及交换容量同步升级。

3.2.2　智能网卡

智能网卡通过卸载网络、存储等功能，从主机 CPU 中释放资源，确保高吞吐量和低时延的传输。

在国际上，NVIDIA 提供了多款智能网卡产品，包括 BlueField2 200 Gbit/s、BlueField3 400 Gbit/s、ConnectX-4 100 Gbit/s、ConnectX-5 100 Gbit/s、ConnectX-6 200 Gbit/s、ConnectX-7 400 Gbit/s、ConnectX-8 800 Gbit/s。Marvell 的 Alaska 88X7120 是支持 400 Gbit/s 网络的以太网卡，主要面向超大规模数据中心，Marvell 也发布了 LiquidIO Ⅱ、LiquidIO Ⅲ两款智能网卡。AMD 发布了 Pensando Pollara 400 AI NIC 以满足现代 AI 环境的要求，同时允许客户基于以太网架构。通过解决 AI/ML 模型的特定通信需求，Pollara 400 能够充分利用其 AI 工作负载的潜力，而不会牺牲以太网基础设施的优势。Intel、Napatech 等网络设备制造商也提供多种型号和配置的智能网卡产品。

在国内，云豹智能提供全功能云霄数据处理单元（data processing unit，DPU）网卡，支持裸金属、虚拟机和容器服务资源一体化，具备热插拔、热迁移和热升级能力，并提供软硬件卸载和加速功能。云脉芯联发布了全自研高性能网络互联芯片 YSA-100，以及 metaScale、metaConnect 等系列智能网卡。另外，华为和中兴也有多款智能网卡产品。星云智联、大禹智芯、益思芯等也相继研发和生产了自己的智能网卡产品。

3.2.3　光模块

作为数据中心内部互联的关键器件，光模块性能直接影响智算集群的整体效率和稳定性。近年来，随着智算集群的快速发展，集群高性能计算和低时延通信的需求对光模块提出了更高的要求，光模块技术也取得了显著进步，主要体现在高速率、低功耗、高密度和智能化 4 个方面。

（1）高速率。随着智能计算集群内部数据交换需求的指数级增长，光模块技术持续向高速率演进：当前，400 Gbit/s 光模块已实现规模化应用，800 Gbit/s 进入商用部署阶段，1.6 Tbit/s 模块正处于研发验证阶段并于 2025 年规模量产。技术突破主要依托高阶调制技术（如 PAM4、相干调制等技术提升单通道速率[28]）、多通道并行架构

（如通过通道数从4路扩展至8路实现带宽倍增[29]）以及磷化铟、硅光子等新型半导体材料的器件性能优化[30-31]，共同推动光互联技术的迭代升级，为智能计算集群构建超高吞吐、超低时延的网络基础设施奠定核心基础。

（2）低功耗。针对智能计算集群规模扩张带来的光模块能耗挑战，低功耗技术研究聚焦于芯片架构优化、封装工艺升级与智能功耗管理三大方向：通过鳍式场效应晶体管（fin field-effect transistor，FinFET）、全耗尽绝缘体上硅（fully depleted silicon-on-insulator，FD-SOI）等先进半导体工艺降低核心芯片功耗[32]；采用铜柱凸点、硅通孔等三维封装技术提升热传导效率以减少封装层能耗[31]；结合自适应均衡和低功耗模式等智能管理策略实现流量动态驱动的能耗调节[33]。这些技术突破使光模块单位比特功耗显著下降，为超大规模智算集群的绿色可持续发展提供了关键支撑。

（3）高密度。智算集群对空间利用率要求较高，需要光模块具备更高的集成度。高密度光模块的实现主要依赖于小型化封装、共封装光学（co-packaged optics，CPO）、板载光学（on-board optics，OBO）等技术：采用更小的封装尺寸，提高单位面积内的端口密度[34-35]；将光模块与交换机芯片封装在一起，缩短信号传输距离，提高集成度[36]；将光学器件直接集成到印制电路板上，进一步缩小体积，提高密度[40-41]。

（4）智能化。随着智算集群规模的扩大和复杂度的提高，光模块的智能化管理日益重要。光模块智能化主要通过多参数状态监测、故障诊断和性能优化等技术，实现运行状态的深度可观测、故障隐患的精准定位和传输性能的动态调优：基于多参数实时监测光模块的工作状态，包括温度、电压、光功率等，构建故障预警系统，结合嵌入式诊断算法对光模块的故障诊断与定位[39]；同时，根据网络流量和业务需求，动态调整光模块的工作参数，优化网络性能[40-41]。

3.3 通信协议与传输优化

主要技术途径包括传输协议、负载均衡、拥塞控制、在网计算和集合通信库优化等多个方面。

3.3.1 传输协议

目前主流的组网协议/传输协议主要有3种：IB、RoCE、互联网广域远程直接内存访问协议（Internet wide area RDMA protocol，iWARP）。

IB协议支持RDMA，由InfiniBand贸易协会（InfiniBand Trade Association，IBTA）提出[42-43]，允许数据直接在服务器之间传输，无须CPU参与，是一种为高性能计算和数据中心设计的网络通信标准，以其低时延、高带宽、高效能的RDMA特性、强大的扩展性和灵活性，在大规模AI模型训练中提供了快速、稳定且可靠的数据传输能力。然而，IB是一种专为RDMA设计的网络，从硬件级别保证可靠传输，但成本高昂且生态封闭，搭建基于IB技术的RDMA网络需要专用的IB网卡和IB交换机。NVIDIA、百度、零一万物等均采用IB协议构建具有万卡规模的智算集群网络。

RoCE是一种基于以太网的RDMA协议，由IBTA提出[44]，通过以太网实现高性能、低时延的数据传输，需要服务器安装RoCE网卡。其依赖于无损以太网传输，要求网络中所有设备支持二层服务质量，如优先级流量控制（priority flow control，PFC），以确保数据传输的可靠性。RoCE协议分为两个版本，RoCEv1和RoCEv2。RoCEv1使用以太网的L2层进行数据传输，并依赖于数据中心桥接（data center bridging，DCB）技术来确保流量的优先级和无损传输，限制在同一个虚拟局域网（virtual local area network，VLAN）内通信，而RoCEv2克服了这一限制，通过在数据包中包含IP和UDP标头，实现了跨L2和L3网络的通信。Meta已采用RoCE协议建设了24K规模的集群网络。

iWARP[44] 是一种基于 TCP/IP 协议栈的 RDMA 技术，由互联网工程任务组（The Internet Engineering Task Force，IETF）标准定义，允许在标准的以太网基础设施上实现高效、低时延的数据通信，服务器须安装支持 iWARP 的网卡。iWARP 的优势在于在不改变现有 TCP/IP 网络架构的情况下，通过 RDMA 技术提高数据传输效率。目前，Intel、Chelsio、NVIDIA、Marvell 等均有实施 iWARP 技术。

iWARP 和 RoCE 是基于以太网的 RDMA 技术，iWARP 基于 TCP/IP 协议，RoCE 在以太网链路层或 UDP/IP 层实现。IB 是一种专为 RDMA 设计的专用网络，提供相对较高的性能，但成本也相对较高。选择哪种技术取决于具体的应用需求、成本预算和现有网络基础设施。

3.3.2　负载均衡

负载均衡[46]是网络领域的经典问题。由于大象流或局部负载不均引发的网络拥塞在数据中心网络中普遍存在，而在 AI 网络环境下这一问题尤为显著。AI 大模型训练流量具有低熵的性质，对网络基于流哈希的负载均衡机制“并不友好”，容易造成局部热点，产生拥塞。目前，解决大模型训练负载不均衡的方式主要有以下 3 种。

（1）静态负载均衡

传统的负载均衡方式依赖于预先设定的规则或配置，这些规则一旦设定不会轻易改变，适用于流量模式相对稳定的情况，但对于变化迅速的网络应用不够灵活。

等价多路径（equal-cost multi-path，ECMP）是一种经典静态负载均衡技术[78]，通过使用一种哈希函数决定将数据包发送到哪个等价路径。它允许数据包在多个具有相同成本的路径之间进行均衡分配，以优化网络利用率、提高吞吐量和冗余性，同时降低单条链路过载的风险。然而，由于大模型训练任务流量具有低熵的性质，采用 ECMP 路由算法性能不佳，这种不完善的负载均衡将导致哈希冲突的概率变高。Meta[6]采用固定路径、E-ECMP、基于 QP-Scaling 的 ECMP 来解决大模型训练任务中 ECMP 导致的哈希问题，但哈希的潜在概率性质是这些路由方案的一个持续缺点。DeepSeek[12]选择静态路由策略，将存储流量均匀分散到 Leaf→Spine 链路中。阿里云 HPN[13]采用双平面设计，通过部署主机–交换机协作系统，运用转载准确路径控制（reprint accurate path control，RePaC）技术确保主机直接获取每个交换机的精确哈希结果，找到所有不相交路径及其对应的五元组（源 IP 地址、目的 IP 地址、协议号、源端口号、目的端口号），并在集合通信库中平衡它们的负载，据此建立 RDMA 连接，从而缓解哈希冲突。字节跳动 MegaScale[15]优化了网络拓扑，并调度网络流量以减少 ECMP 哈希冲突。在机架顶部交换机（top-of-rack，ToR）级别分裂 400 Gbit/s 下联端口为两个 200 Gbit/s 下联端口，由于每个上行链路的带宽是下行链路带宽的两倍，因此冲突概率会降低。其次，服务器上的 8 个 200 Gbit/s 网卡以多轨方式连接到 8 个不同的交换机。同一组 ToR 交换机连接的 GPU 服务器数量可以达到 64 台，通过策略性调度数据在同一个 ToR 交换机下运行，可以显著减少通信所需要的交换机跳数，进一步降低 ECMP 哈希冲突的概率。

（2）动态负载均衡

引入实时监测和自适应调整的能力，根据当前的网络状况（如链路负载、时延等）动态地重新分配流量，显著提高了响应速度和资源利用率。

Meta TE[6]根据实时工作负载和拓扑输入动态优化路由，通过设计控制平面根据实时工作负载和拓扑输入动态优化路由。百度云联合亲和性调度和动态负载均衡（dynamic load balancing，DLB）来解决哈希冲突。前者通过优化任务调度减少不必要的流量传输，允许通过指定亲和性规

则来控制Pod与节点进行亲和匹配。这种调度策略有助于减少流量上送到Leaf交换机，从而降低交换机哈希冲突的概率。后者通过动态负载均衡技术在物理层面上分散流量，突破传统静态哈希机理限制，通过引入时间戳、实时负载度量（如端口带宽负载、队列大小）因子，在时间、带宽空间两个维度优化负载均衡效果，提供动态、智能的哈希机制，两者结合使用显著提高网络效率和性能。

（3）AI驱动的负载均衡（AI-driven load balancing，AILB）

目前的最新趋势是利用AI和机器学习算法来预测流量模式、优化路径选择，并自动化决策过程。通过持续学习网络行为，AILB可以更加智能地应对复杂多变的环境，提供最优的流量管理和故障恢复策略。华为推出了网络智能负载均衡（network smart load balancing，NSLB）算法，当NSLB与神经网络处理单元（neural processing unit，NPU）协同工作时，控制器从全局主动获取或解析AI流量通信关系，并统一计算路径和下发配置，这样可以消除整个网络中的链路冲突。当NSLB与GPU配合使用时，网络可以主动检测拥塞并自动切换路径，实现全网负载均衡。

3.3.3 拥塞控制

集群网络处理拥塞[63-64]的经典技术主要有3种：PFC、ECN与DCQCN。PFC[47,49]由IEEE 802.1Qbb[50-51]标准定义，允许网络设备根据流量的优先级进行拥塞控制。PFC通过在交换机之间发送PAUSE帧来控制流量，从而避免拥塞。ECN机制最早在RFC 3168中提出，并在2001年正式标准化[52]。ECN通过在IP头部添加显式拥塞标记，允许网络设备在检测到拥塞时通知发送方，从而实现更有效的拥塞控制。DCQCN数据中心量化拥塞通知[47]是ECN和PFC的组合，可支持端到端无损以太网。DCQCN的设计理念是在拥塞时通过ECN让发送端降低传输速率，从而尽量避免触发PFC。DCQCN需要考虑两个关键问题：一是确保PFC不会太早触发，即先使用ECN发送拥塞反馈使流量变慢；二是确保PFC不会太晚触发，即拥塞较严重产生缓冲区溢出进而出现丢包。因此，在DCQCN算法的使用过程中，根据实际业务调整ECN阈值是非常关键且具有挑战性的议题。

在已有文献中，Meta[6]抛弃DCQCN算法，依赖PFC进行流量控制。首先，通过集合通信库接收方驱动的流量准入与交换机深度缓冲区来进行拥塞控制，接收方驱动的流量准入控制利用NCCL集体库，通过调整通道数量和通道缓冲区大小限制网络流量，特别是在拥塞开始发生时；其次，利用具有深缓冲区的集群训练交换机来吸收短暂的拥塞，以减少头部阻塞。DeepSeek[12]采用IB的服务级别（service level，SL）技术[53-54]，针对4种不同类型的流量HFReduce通信、NCCL通信、3FS存储流量和其他流量，在节点之间建立连接时分配不同的SL值，并将SL映射到IB物理队列虚拟通道。采用虚拟通道可确保不同通道中的流量不会相互干扰，并配置它们的比例从而实现流量隔离，防止堆场阻塞[55]和不同的流量冲突引起的网络拥塞。阿里云[13]HPN实现了一种应用层的负载均衡方案，通过在集合通信库中维护一个计数器记录当前活动工作队列元素的总字节数，作为揭示当前连接拥塞状态的指示。计数器显示当前连接的拥塞状态，拥塞的连接会减慢工作队列的耗尽速度。因此，该方案选择计数器最小的连接来发送消息，从而在多个RDMA连接之间实现负载均衡。字节跳动MegaScale[15]结合Swift和DCQCN的原则，通过精确测量显式拥塞通知和ECN的快速拥塞响应能力，显著提高吞吐量并最小化与PFC相关的拥塞。MegaScale调整NCCL中的参数，以控制重传计时器和重试次

数，在链路闪烁时快速恢复。百度云基于IB的集群网络[17]通过ECN机制、拥塞控制参数调整、拥塞行为观测以及亲和性调度与DLB动态负载均衡的结合，实现了IB网络中的有效拥塞控制。

因此，由于AI大模型训练产生新的流量模式，采用传统数据中心DCQCN算法调节ECN参数设计是具有挑战的，需要结合任务的属性进行灵活设计。然而，得益于大模型训练流量的预测性，设计控制器进行预先防止的方法具有借鉴意义，结合网络实时拓扑选择最优路由，同时对新的路由进行实时评估，预测拥塞并修改路由；借助集合通信库设计将拥塞算法从被动转化为主动拥塞控制算法是长远且有效的方式，也是未来比较重要的发展方向。

3.3.4　在网计算

在网计算通过在网络交换机内部执行部分计算任务，有效减少了数据传输量和通信开销，减轻了GPU的负担，从而提高GPU训练效率，特别是在处理大规模集合通信操作时，对于需要处理和传输大量数据的AI训练任务至关重要。

以在网计算为核心的分布式机器学习优化技术近年来成为学术界的研究热点。文献[56]提出的PANAMA框架采用定制化的硬件加速器支持浮点梯度聚合，并结合多树负载均衡和轻量级拥塞控制协议，在共享集群中实现高吞吐和低时延的梯度同步，其现场可编程门阵列（field programmable gate array，FPGA）原型验证了在10 Gbit/s链路上无损聚合的可行性。文献[57]提出了一种针对现代多机架、多作业设置的网络内聚合传输协议（aggregation transport protocol，ATP）。通过可编程交换机实现动态、尽力而为的聚合策略，支持跨机架的多级聚合和哈希冲突规避机制，在资源竞争环境下仍能提升38%~66%的训练吞吐量。文献[58]提出SwitchML，通过聚合来自网络中多个工作节点的模型更新来减少交换数据量，与终端主机协议和ML框架共同设计了交换处理，为许多真实世界的基准模型加速高达5.5倍的训练。文献[59]提出了一种面向All-Reduce操作的路由器微架构，通过集成RDMA协议、高效内存管理和流水线优化，显著提升了大规模分布式训练的聚合吞吐量。文献[60]系统梳理了数据中心网络中网内聚合技术，从可编程交换机、中间设备和新型交换机架构3类硬件实现方法展开分析，并探讨了聚合算法设计、性能对比及未来挑战。

在网计算技术在产业界也有了相对成熟且落地的工程实践案例。国际上，可伸缩的分层聚合和缩减协议（scalable hierarchical aggregation and reduction protocol，SHARP）是一种由NVIDIA开发的网络通信算法[61]，用于优化和加速高性能计算和AI场景中的集合通信操作。通过将深度学习训练（如All-Reduce梯度同步）和科学计算中的集合操作从CPU/GPU卸载至支持智能计算的网络交换机硬件，减少数据在网络中的传输次数及穿越网络的数据量，显著降低了集合通信操作的时间。实现SHARP算法需要网络硬件（尤其是交换机芯片）的支持。这些硬件需要具备处理和聚合数据的能力，通常通过在硬件中嵌入专门的处理单元来实现。在国内，华为NetReduce[62]是一种新型的RDMA兼容的网络内归约架构，旨在加速分布式深度神经网络训练。与传统设计相比，NetReduce通过在网络内进行归约操作来提高分布式训练的效率。

上述工作表明，在网计算不仅通过硬件加速优化了任务性能，而且通过灵活的协议设计和资源管理，为多租户集群提供了高效的网络资源共享能力，为未来智算集群的规模化部署奠定了基础。

3.3.5　集合通信库——优化集合通信的效率

使用并行计算方案来有效利用多个计算资

源，提高资源利用率可满足大模型训练的高计算和大存储需求。目前，并行方案主要有数据并行、模型并行、流水并行、张量并行、专家并行、序列并行等。这些并行方案需要通过集合通信库中的通信原语（如All-Reduce、Broadcast、All-Gather等）实现，使大模型训练能够在多GPU、多节点的分布式环境中高效进行。

NCCL是NVIDIA推出的针对GPU集群的开源通信库，用于在多个NVIDIA GPU之间实现快速的数据传输和协同计算，提供All-Gather、All-Reduce、Broadcast、Reduce、Reduce-Scatter等用于参数同步、梯度聚合等操作，已广泛应用于TensorFlow、PyTorch等深度学习框架。阿里推出ACCL集合通信库，采用了Halving-Doubling等算法来实现All-Reduce操作，在特定网络拓扑下能够实现全局无拥塞，适用于大规模分布式训练场景，提高训练效率。华为推出HCCL是基于昇腾AI处理器的高性能集合通信库，提供单机多卡及多机多卡间的集合通信能力，可支持多种通信算法，如Ring、Mesh、Halving-Doubling等，并支持基于昇腾片间通信服务（AiAI Chip connect service，HCCS）、RoCE和PCIe等链路/协议实现集合通信。BCCL是百度研制的高性能集合通信库，基于NCCL进行了功能扩展和能力增强，针对大模型训练场景在可观测性、故障诊断、稳定性等方面进行优化，进一步提升集合通信库的可运维能力。TCCL是针对腾讯云星脉网络架构的高性能定制加速通信库，可为AI大模型训练提供高效的网络通信性能，同时具备网络故障快速感知与自愈的智能运维能力。TCCL基于NCCL代码做了扩展优化，兼容NCCL的功能与使用方法，支持双网口动态聚合优化、全局哈希路由、拓扑亲和性流量调度、最小化流量绕行等特性。对比NCCL方案，TCCL在大模型训练场景下预计可提升约50%带宽利用率。

对集合通信库的优化可以提高大模型训练过程中网络通信的性能，国内各大厂商已经陆续在NCCL基础上展开研发，这些集合通信库通过提供高效的数据同步和规约操作，使得在分布式环境中进行深度学习训练变得更加高效和可行。

3.4 网络故障检测与恢复

3.4.1 网络状态监控

随着网络复杂度和规模的不断增长，网络状态监测在确保网络健康、优化性能、及时响应故障方面变得越来越重要。网络状态监测技术主要分为3类：主动测量、被动测量和混合测量。主动测量通过向网络中主动传送探测分组，并根据探测分组受网络影响而发生的特性变化来分析网络行为。常见的主动测量技术包括Ping、跟踪路由、带内遥测（in-band network telemetry，INT）等。被动测量通过捕获流经测量点的分组来测量网络状态、流量特征和效能自变量。常见的被动测量技术包括流量采样（如NetFlow、sFlow）和深度包检测。混合测量技术是主动测量与被动测量相结合的一种方法。通过结合两者的优势，混合测量能够在提供精确网络性能数据的同时，避免主动测量对网络的干扰。典型的混合测量应用包括网络性能监控平台（如SolarWinds、Nagios），通过结合主动与被动测量方法，提供详细的网络状态监测。

网络监测系统通常需要具备高效的实时数据采集、存储、展示和报警功能。Prometheus[71]是一个开源的监控系统，用于收集和存储时序数据，广泛应用于系统和服务监控，其核心功能是从配置的网络设备或服务器中拉取各种性能指标数据，并将数据存储在时间序列数据库中。Prometheus的数据模型能够灵活地采集来自不同网络设备、路由器、交换机及服务器的实时性能指标，如带宽利用率、时延、丢包率等，通过定期轮询网络设备的指标端点，获取相关

性能数据，并提供了查询语言PromQL，使用户方便地对网络状态进行实时查询和分析。Grafana[71]是一款开源数据可视化工具，可与Prometheus等监控系统无缝集成，帮助用户以图形化的方式展示数据。Grafana支持丰富的可视化图表，支持用户创建自定义的仪表盘，展示网络监测数据，如流量、时延等，实现实时监控和分析，帮助管理员直观地了解网络健康状况和性能趋势。Grafana还提供报警功能，当某些网络性能指标超过预设阈值时，自动触发警报，并通过邮件、Slack等通知用户，网络管理员可以第一时间获得网络故障或性能问题的预警，及时采取措施。

Meta[6]采用带内遥测对网络中的交换机、网卡、PCIe、GPU进行监控和检测，定义3个重要的计数器来识别网络问题，包括乱序、链路抖动计数器、本地确认超时等，并在网络中设置了PFC看门狗、缓冲区阈值和拥塞丢弃、可达性检测等保护机制。百度云[17]通过智算网络秒级监控运维平台，使用Netconf、Telemetry等技术实现对智算网络的流量监控、端口top *N*、反压报文PFC指标、拥塞通告报文等的监控。字节跳动MegaScale[15]开发的诊断工具深入监控系统组件和事件，帮助识别复杂问题的根源，包括心跳消息、实时异常检测、故障定位和恢复过程的优化。通过心跳消息，系统能够实时检测到执行器的基本信息和训练进程的状态，从而实现实时异常检测和预警。

3.4.2　网络故障定位

传统的故障定位方法多基于网络拓扑结构，通过分析节点间的连接性和性能数据，来确定故障发生的位置。常见的故障定位方法包括以下5类。

（1）基于日志分析的方法：网络设备通常会记录日志信息，故障发生时，这些日志提供关于故障的线索。通过分析日志信息，尤其是结合时间戳，可以快速定位到发生故障的设备或链路[72]。

（2）基于路径跟踪还原的方法：路径跟踪技术通过不断探测数据包的路径，如采用ICMP协议的Ping命令或Traceroute工具，逐步跟踪数据包的转发路径，一旦检测到故障点，就可以快速定位到故障发生的环节[72]。在大规模训练任务中，涉及多对多的GPU通信，故障链路长，需要关联分析多种网元。华为云通过自研的全链路诊断系统，结合全景图配置数据库，快速计算训练任务涉及的任意一台源和一台目的主机的流量路径，路径涵盖服务器、GPU卡、端口、链路、单板、交换机等所有网元，实现故障快速定界。

（3）基于流量分析的方法：通过对流量数据的分析，监控流量的异常变化或丢包情况，结合网络拓扑数据，可有效判断流量受阻的设备或路径，从而进行故障定位[72]。

（4）基于机器学习的方法：随着AI技术的发展，基于机器学习的网络故障定位方法逐渐兴起。通过训练模型对历史数据进行学习，识别出不同类型故障的特征，结合实时数据进行故障预测和定位，该方法可在一定程度上减少人为干预，提高故障诊断的准确性和效率[73]。如在亚健康光模块故障定位中，利用AI算法结合光模块的历史故障特征、日志特征、指标特征等多维度数据，进行故障评分和寿命预测，提前预防问题发生。

（5）基于多监测手段的综合监控与诊断：通过集成多种监控信号源，如白盒监控、黑盒监控、流量监控、传输监控等，实现故障的综合定位[74]。

在已有文献中，Meta[5]采用NCCLX（NVIDIA NCCL库的一个分支）与PyTorch紧密合作，提高了故障检测和定位的速度和准确性，允许PyTorch访问NCCLX的内部状态并跟踪相关信息。

当检测到NVLink故障导致的停滞时，系统会监控通信库的状态并自动超时。字节跳动MegaScale[15]开发了一个3D并行训练可视化工具，显示了数据依赖性和不同通信操作的逻辑拓扑，通过性能分析工具记录每个机器排名在运行时的关键代码段执行时间，并通过热图和分布式视图的事件时间线提供不同的分析视角。

3.4.3 网络故障恢复

网络故障恢复技术是保障网络高可用性和业务连续性的重要手段。该技术不仅能有效识别和隔离故障，还能在故障发生后尽快恢复网络服务，减少故障对业务的影响。网络故障快速隔离与恢复的技术主要包含以下4个方面。

（1）状态保存点（Checkpoint）的状态保存与恢复技术：在网络的不同状态或阶段定期保存系统的关键状态信息，以便在发生故障时迅速恢复到之前保存的健康状态。Checkpoint通常保存系统的配置、连接状态、路由信息等关键数据。当网络发生故障时，通过恢复到最近的Checkpoint，系统可以避免从零开始重新建立连接，减少人为干预和复杂恢复操作的需求，显著加快故障恢复速度[75]。

（2）快速故障隔离技术：在故障发生时，迅速识别并隔离故障源，避免故障进一步蔓延，包括网络拓扑感知技术，通过实时监控网络拓扑，识别出故障发生的位置，并根据网络拓扑迅速调整数据流，避免故障影响其他网络区域；故障预测和预警系统利用机器学习和大数据分析技术，通过对网络流量、设备状态和性能指标的实时监控，预测潜在的故障风险，并提前触发预警，进行预防性隔离[77]。

（3）快速重路由和负载均衡技术：该技术是网络故障恢复的常见手段，在故障发生时，快速重路由技术通过动态调整路由策略，确保流量能够绕过故障节点或链路，恢复网络连接。同时，负载均衡技术通过将流量智能分配到多个路径或设备上，确保在某一路径发生故障时，其他路径或设备能够分担流量，保证服务的连续性[76]。

（4）动态流量控制和服务质量技术：帮助网络在发生故障时，优先保证关键业务的流量和带宽。如网络可以基于业务的优先级和实时性能需求，对流量进行优先级排序，在出现故障时对重要流量进行优先转发。这样，虽然网络受到故障影响，关键应用和服务仍能得到保障[77]。

在已有文献中，阿里云HPN[13]在网络发生故障时，主机需要从ToR交换机获取新的ECMP组，然后重新计算不相交路径，而不是在全局控制器中维护来自不同层级的ECMP组。这不仅简化了故障恢复过程，也减少了路径重新计算的复杂性。字节跳动MegaScale系统[15]检测到异常状态或超时未收到心跳时，驱动进程会触发故障恢复程序，包括挂起所有执行器上的正在进行的训练任务，并运行一系列自检诊断测试。为了减少训练中断，MegaScale优化了检查点和恢复过程，将从最新检查点恢复训练的时间最小化。

3.4.4 网络拓扑结构冗余设计

在网络拓扑结构设计中，通过增加冗余链路可有效提高网络运行的可靠性，如双上联拓扑结构可为超大规模智算集群提供高鲁棒性支撑。阿里云HPN[13]通过双平面设计、双上联策略等技术手段，有效提升了大模型集群网络的稳定性和可靠性。双上联策略和双平面架构的结合，使得系统在面对网络故障时能够保持稳定运行。一旦上行链路或对应交换机故障，流量将切换至另一端口提供服务，训练任务不会中断，从而避免了单点故障导致的整体网络服务中断。

本文针对各个技术领域当前主要的技术总结了挑战与技术途径，见表3。

表3　挑战与技术途径

细分技术途径	归属技术领域	“提升智算集群网络互联和通信能力”的主要技术途径	“提高智算集群网络传输效率”的主要技术途径	“增强智算集群网络运行的可用性”的主要技术途径
网络拓扑设计	网络架构	Fat-Tree、Spine-Leaf、Dragonfly、Slimfly	—	—
组网形态	网络架构	单轨、轨道优化、双平面[13]	轨道优化[13]	双上联、双平面[13]
GPU互联方式	网络架构	PCIe技术[26]、NVLink[26]、UALink[27]	—	—
高性能网络交换机	网络设备	H3C、云尖、锐捷、中兴	—	—
智能网卡	网络设备	云豹、云脉芯联	—	—
集合通信库	网络架构、通信协议	FlagCX、BCCL、HiCCL、IHCCM	NCCL、ACCL、HCCL、BCCL、TCCL	—
传输协议	通信协议	IB[42]、RoCE[44]、iWARP[45]	IB[42]、RoCE[44]、iWARP[45]	—
负载均衡	通信协议	—	ECMP、E-ECMP[6]、集中流量工程[6]	ECMP、E-ECMP[6]、集中流量工程[6]
拥塞控制	通信协议	—	PFC[50]、ECN、DCQCN[47]、亲和性调度[17]、集中控制器	PFC[50]、ECN、DCQCN[47]、亲和性调度[17]、集中控制器
在网计算	通信协议	—	SHARP[61]、NetReduce[62]、PANAMA[56]、ATP[57]、SwitchML[58]	—
网络状态监控	故障检测与恢复	—	—	Prometheus[71]、Grafana[71]、KubeNurse、Ganglia
网络故障定位	故障检测与恢复	—	—	基于日志分析、路径跟踪还原、流量分析、机器学习[9,72-73]
网络故障恢复	故障检测与恢复	—	—	Checkpoint技术[75]、快速故障隔离、快速重路由与负载均衡[76]、动态流量控制与QoS[77]

4　技术发展趋势

4.1　总体发展趋势

当前基于国产算力的超大规模异构智算集群网络正朝着高性能、智能化、绿色节能和高可靠性的方向快速发展，通过引入先进的软硬件技术（如400 Gbit/s及以上速率的交换机、智能网卡、RDMA/RoCE协议等）、优化网络拓扑结构、采用AI驱动的运维工具来提升数据传输效率和自动化管理水平；同时，推动多类型算力资源的有效整合，实现更为广泛的互操作性，并积极探索在网计算、负载均衡、拥塞控制、故障检测恢复等技术机制，确保即使在网络复杂度不断增加的情况下仍能维持高效稳定的运行和服务质量，支持日益增长的大模型训练及其他复杂计算任务的需求。

4.2　各技术领域发展趋势

4.2.1　网络架构

（1）网络拓扑结构

当前国产超大规模异构智算集群网络普遍采用Fat-Tree网络拓扑，其通过分层交换机架构实现非阻塞数据传输、高二分带宽和弹性扩展能力，兼具低时延、高吞吐与强容错性，支撑多样化AI负载需求。Dragonfly与Torus的规则网格在特定场景（如集合通信）具备理论优势，但受限于扩展复杂性（须物理重构布线）、应用适配局限及生态成熟度不足，未形成规模化应用。近年突破性技术Slim Fly拓扑[23]通过将网络直径压缩至2，实现了比Clos结构降低25%~30%成本/功耗、50%时延，其首个大规模部署案例[24]验证了物理布局可行性，推动新型低直径网络拓扑的研究，正逐步打破传统Clos的垄断格局。此外，为

解决固定网络拓扑结构灵活性不足的问题，可重构数据中心网络正通过链路、层、拓扑等技术层次上的重构来灵活配置数据中心拓扑结构，以适应大模型训练需求的不断变化，提升网络基础设施的建设和降低使用成本。

（2）网络互联方式

基于CPU的网络互联方式目前较为普遍，可利用高性能CPU来处理和转发数据包，并通过高速网络接口实现节点间的连接。围绕CPU构建的网络互联方式拥有成熟的软件栈和技术生态，支持广泛的协议和工具，便于管理和维护，能有效避免不同GPU间通信库差异带来的潜在问题，虽然会增加CPU与GPU之间数据复制的开销，但这种权衡显著换取了异构混训系统的稳定性和兼容性。然而，基于GPU的网络互联方式能够提供更高的传输效率，支持高带宽内存直接访问特性，使该互联方式能够在数据传输过程中直接在GPU之间交换数据，减少了通过CPU中转所带来的时延和开销。未来，越来越多的国产算力GPU将能够支持远程直接内存存取RDMA，允许数据绕过主机CPU直接在GPU间传递，极大地提升了传输速率和降低了系统负担。随着国内技术的不断发展，基于GPU的网络互联方式正逐渐获得更多运用。

（3）集合通信库

实现异构算力集群中集合通信的通用性，需要从标准化接口、自适应优化和跨架构支持等多方面入手。智源研究院的FlagCX、百度的BCCL、斯坦福大学HiCCL[70]等项目已经在这方面取得了重要进展，但未来仍须进一步推动标准化和开源合作，以满足日益增长的异构算力需求。

4.2.2 网络设备

（1）高性能交换机

新一代交换机将朝着更高的端口密度和更快的数据传输速率发展，如400 Gbit/s甚至800 Gbit/s端口逐渐普及，满足了大规模数据流量的需求。交换容量也正朝着更强大的数据交换能力持续演进，逐渐从25.6 Tbit/s跃迁至51.2 Tbit/s，甚至更高的102.4 Tbit/s。面向自主可控需求，全国产化网络交换机的研制工作也在进行中。同时，传统基于电交换的网络架构面临着带宽瓶颈、时延增加及能耗过高的挑战，业界正逐渐转向光学技术，基于可重构光交换机（optical circuit switching，OCS）来构建更加灵活且高性能的网络基础设施将成为一种可行的潜在方案，可进一步增强网络架构对未来智能计算任务的适应性。

（2）智能网卡

智能网卡和DPU的技术发展趋于智能化，卸载功能也进一步增强，不仅具备传统网卡的基本功能，还集成了额外的计算资源，用于执行诸如在网计算、存储加速、拥塞控制等功能，减轻主机CPU负担。越来越多的网卡开始支持RDMA和RoCE协议[44]，显著降低了通信时延并提高吞吐量。通过P4语言或其他类似工具实现的可编程网卡，获得网络可编程能力支持，允许用户根据具体应用场景自定义数据包处理逻辑，增强了灵活性和适应性。

（3）光模块

光模块在低功耗与长距离传输方面的能力将进一步强化，特别是随着硅光子技术和相干光学的进步，光模块能在保持低功耗的同时实现更远距离的数据传输。同时，光模块将具备更好的多速率兼容性，支持多种传输速率（如100 Gbit/s、200 Gbit/s、400 Gbit/s等），并且可通过软件配置灵活切换，增强系统的兼容性和扩展性。为了适应更高密度的部署需求，光模块不断向小型化方向发展；同时，为了便于维护和升级，采用热插拔设计，以提高操作便利性和可靠性。

4.2.3 通信协议与传输

（1）传输协议

RDMA及其变种如RoCEv2持续得到优化和支持，通过硬件卸载机制减少CPU负载，提供

低时延、高带宽的数据传输。RDMA 通信协议也将进一步优化 QP 机制，采用非连接的可靠传输或动态连接池资源共享机制。传输协议所采用的重传机制，已从 Go back *N* 逐渐转向选择性重传。同时，持续探索新的传输协议，以及针对特定应用场景定制的轻量级协议也是未来发展的趋势。

（2）负载均衡

负载均衡机制调度的粒度正在从传统的逐流[46]调度向着更精细化的逐子流[6]、逐包[46]发展。流量调度的方法从传统的依赖于预先设定的规则或配置的静态方式[46]，向着动态、自适应[6,46]（根据实时监测到的网络流量负载情况自动流量传输路径，实现自适应调整）和基于 AI 生成式的方式演进，提供智能流量调度能力，利用机器学习算法实现动态流量预测和路径选择，确保整个网络资源的最佳利用，避免局部热点形成。

（3）拥塞控制

网络侧的拥塞控制方式主要基于全局或逐跳的队列调度实现控制和精细反压，而端侧则通过实时精细化感知网络状态信息，如时延、网络队列长度等变化，进行速率调整。同时，控制机制也逐渐从端网分离向端网协同控制过渡和发展；计算任务的分配方式也逐渐纳入拥塞控制机制的管理范畴，与网络通信的调度机制进行协同，实现算网融合发展，促进大模型训练业务流量的更高效传输。越来越多的研究也借助集合通信的操作实现拥塞控制，细粒度调控机制进一步普及，对不同类型的流量实施差异化的拥塞控制策略。智能化决策手段也将获得更为广泛的应用，结合深度学习模型进行实时数据分析，预测未来可能出现的拥塞点，并采取预防措施，提前重定向部分流量。

（4）在网计算

在网络设备中集成计算能力将成为主流技术发展趋势，在各类网络设备中嵌入更多的计算资源，允许在网络层直接处理某些数据运算，减少数据传输成本。在大模型训练中，在网计算助力分布式训练来提高效率，如通过在网内进行梯度聚合，减少数据传输带宽需求。

（5）集合通信库

集合通信库的算法实现将获得进一步的优化，开发自适应拓扑感知算法，动态优化 All-Reduce、Broadcast 等集合通信模式，进一步减少通信次数和时间复杂度。通信库对于异步通信的支持也将增强，通过对异步通信模式的支持，允许任务在等待结果的同时继续执行其他工作，从而提高系统的并发度和响应速度。考虑国产异构算力的使用需求，集合通信库跨硬件平台兼容性方面亟待提升，需构建异构计算统一通信框架，确保集合通信库能够在多种硬件平台上无缝运行，同时保持良好的性能表现。开源生态方面，推进兼容 NCCL 的自主集合通信库研发，深度集成梯度压缩、通信－计算联合优化等创新技术。

4.2.4　网络故障检测与恢复

（1）网络状态监控

网络状态监控正向超细粒度主动感知演进，基于智能网卡的内联式数据采集 INT 逐步取代传统探针，结合光子晶体光纤传感器，可实现亚微秒级时延测量与高带宽链路的无侵入式监控，实现对网络性能指标的实时采集。同时融合物理拓扑、逻辑通信与训练任务数据构建三维健康图谱，并基于数字孪生平台通过强化学习预判风险，推动监控模式从“被动响应”向“预测干预”转型，提升大规模训练场景的实时性与全局可视性。

（2）网络故障定位与检测

引入 AI 驱动的自动化故障诊断工具，自动解析日志文件、配置信息和其他相关数据，快速确定故障原因及位置。同时，突破固定阈值告警范

式，利用无监督学习建立动态流量基线，结合全栈追踪技术实现从GPU显存到光模块的端到端调用链可视化，结合多个因素（如流量模式、设备状态、环境条件等）之间的关联性进行深度挖掘，准确定位复杂网络环境下的故障根源。

（3）网络故障恢复

网络故障恢复和自愈机制迈向意图驱动与硬件加速阶段，通过数字孪生预演恢复策略并利用深度强化学习生成最优路径，如All-Reduce中断时自动切换混合并行模式；结合基于IPv6的段路（segment routing over IPv6，SRv6）/软件定义网络（software-defined network，SDN）实现链路故障的智能重路由与计算任务协同续跑，同时依托存算一体芯片在智能网卡本地执行拥塞控制算法，直接调整流优先级或触发ECN标记，加快故障恢复的进程。

4.3 自主可控需求的开放式探讨

为避免我国在关键技术领域出现被“卡脖子”的困境，智算集群网络面临自主可控要求，需要着力在多个方面系统性地推进技术创新。在交换芯片方面，应加强在芯片设计与制造方面的自主研发能力，攻克低时延和高带宽的设计挑战，以支持更高的数据传输速率和更大的交换容量。在网络架构方面，探索适合国产智算集群的新型网络架构，并针对异构国产算力混合组网和训练的迫切需求，依据计算能力与网络能力相匹配的原则进行拓扑方案优化。如借鉴阿里云提出的双平面+双上联架构或其他创新性设计方案，旨在提高网络的灵活性、可扩展性和容错能力；为应对大语言模型训练对网络带宽的严苛需求，华为提出一种新型AI数据中心网络架构UB-Mesh[79]，在架构设计方面取得重要进展。在网卡方面，开发集成RDMA加速、协议卸载、资源隔离功能的国产智能网卡，加快自主新型端到端拥塞控制机制的功能实现，增强对虚拟化的支持，有效实现云环境下的资源共享与隔离。如云豹、云脉芯联等国内厂商已研制具备部分功能的网卡产品。在传输协议方面，研发更适合国产智算中心需求的新兴传输协议。在集合通信库方面，构建面向国产异构算力的统一集合通信库，支持多种国产算力资源之间的协同工作和混合训练，并持续优化集合通信库，加速模型收敛速度，减少模型训练成本。构建智算集群网络标准体系，主导制定智算网络协议标准，参与开放网络基金会（Open Network Foundation，ONF）、开放计算项目（Open Compute Project，OCP）等国际组织。促进产学研用融合，建设国家级智算网络试验场，开展新兴技术的规模化验证。

5 结束语

本文针对当前智算集群建设中面临的依赖进口GPU、成本高昂以及低效的算力资源利用率等问题，强调了构建基于国产计算资源的超大规模智算集群的重要性，以期打破现有硬件供应限制，降低建设运维成本，并提高算力资源利用率。面对国产算力的不成熟状态及其对智算集群网络带来的新挑战，本文深入分析并确定了“如何提升智算集群网络互联能力”“如何提高智算集群网络传输效率”和“如何增强智算集群网络运行的可用性”作为集群网络需要克服的主要障碍，并提出了应对上述重要挑战的总体应对思路，研究涵盖了网络架构设计、网络设备优化、通信协议选择及网络故障检测恢复机制在内的全面技术途径，为促进我国AI基础设施的快速发展提供理论与技术支持。

参考文献：

[1] NVIDIA. NVIDIA DGX. SuperPOD: next generation scalable infrastructure for AI leadership[R]. 2023.

[2] 徐明强. 微软高性能计算服务器[M]. 北京: 人民邮电出版社, 2010.

XU M Q.Windows HPC server:step by step[M]. Beijing: Posts & Telecom Press, 2010.

[3] JAISWAL S, JAIN K, SIMMHAN Y, et al. SageServe: optimizing LLM serving on cloud data centerswith forecast aware auto-scaling[J]. arXiv preprint,2025,arXiv:2502.14617.

[4] COHEN O, SCHAPIRA J Y M, BELKAR S, et al. Routing for large ML models[J].arXiv preprint,2025,arXiv:2503.05324.

[5] DUBEY A, JAUHRI A, PANDEY A, et al. The llama 3 herd of models[J]. arXiv preprint, 2024,arXiv:2407.21783.

[6] GANGIDI A, MIAO R, ZHENG S B, et al. RDMA over Ethernet for distributed training at meta scale[C]//Proceedings of the ACM SIGCOMM 2024 Conference. New York: ACM, 2024: 57-70.

[7] SMITH M S. Challengers are coming for Nvidia's crown: in AI's game of thrones, don't count out the upstarts[J]. IEEE Spectrum, 2024, 61(10): 40-44.

[8] SCHNEIDER I, XU H, BENECKE S, et al. Life-cycle emissions of AI hardware: a cradle-to-grave approach and generational trends[J]. arXiv preprint arXiv:2502.01671, 2025.

[9] HU H, YANG S, ZENG L, et al. US-China trade conflicts and R&D investment: evidence from the BIS entity lists[J]. Humanities and Social Sciences Communications, 2024, 11(1): 829.

[10] LIU A, FENG B, XUE B, et al. DeepSeek-v3 technical report [J]. arXiv preprint,2024,arXiv:2412.19437.

[11] WANG W Y, GHOBADI M, SHAKERI K, et al. Rail-only: a low-cost high-performance network for training LLMs with trillion parameters[C]//Proceedings of the 2024 IEEE Symposium on High-Performance Interconnects (HOTI). Piscataway: IEEE Press, 2024: 1-10.

[12] AN W, BI X, CHEN G T, et al. Fire-flyer AI-HPC: a cost-effective software-hardware co-design for deep learning[C]// Proceedings of the SC24: International Conference for High Performance Computing, Networking, Storage and Analysis. Piscataway: IEEE Press, 2024: 1-23.

[13] QIAN K, XI Y Q, CAO J M, et al. Alibaba HPN: a data center network for large language model training[C]//Proceedings of the ACM SIGCOMM 2024 Conference. New York: ACM Press, 2024: 691-706.

[14] GUO D, YANG D, ZHANG H, et al. Deepseek-R1: incentivizing reasoning capability in LLMs via reinforcement learning[J]. arXiv preprint, 2025, arXiv: 2501. 12948.

[15] JIANG Z, LIN H, ZHONG Y, et al. MegaScale: scaling large language model training to more than 10 000 GPUs[C]//Proceedings of the 21st USENIX Symposium on Networked Systems Design and Implementation. Boston, MA, USA: USENIX Association, 2024: 745-760.

[16] 中国移动通信集团有限公司 . 2024 年面向超万卡集群的新型智算技术白皮书[R]. 2024.
China Mobile. New intelligent computing technology white paper for ultra-large-scale clusters (2024) [R]. 2024.

[17] 百度智能云 . 智算中心网络架构白皮书[R]. 2025.
Baidu AI Cloud. White paper on intelligent computing center network architecture [R]. 2025

[18] CLOS C. A study of non-blocking switching networks[J]. Bell System Technical Journal, 1953, 32(2): 406-424.

[19] LEISERSON C E. Fat-trees: universal networks for hardware-efficient supercomputing[J]. IEEE Transactions on Computers, 1985, C-34(10): 892-901.

[20] AL-FARES M, LOUKISSAS A, VAHDAT A. A scalable, commodity data center network architecture[J]. ACM SIGCOMM Computer Communication Review, 2008, 38(4): 63-74.

[21] Cisco Systems. Cisco data center spine-and-leaf architecture: design overview[R]. 2020

[22] KIM J, DALLY W J, SCOTT S, et al. Technology-driven, highly-scalable dragonfly topology[C]//Proceedings of the 2008 International Symposium on Computer Architecture. Piscataway: IEEE Press, 2008: 77-88.

[23] BESTA M, HOEFLER T. Slim fly: a cost effective low-diameter network topology[C]//Proceedings of the SC '14: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis. Piscataway: IEEE Press, 2014: 348-359.

[24] BLACH N, BESTA M, DE SENSI D, et al. A high-performance design, implementation, deployment, and evaluation of the slim fly network[C]//Proceedings of the 21st USENIX Symposium on Networked Systems Design and Implementation. Boston, MA, USA: USENIX Association, 2024: 1025-1044.

[25] LAKSHMIVARAHAN S, DHALL S K. Ring, torus and hypercube architectures/algorithms for parallel computing[J]. Parallel Computing, 1999, 25(13/14): 1877-1906.

[26] LI A, SONG S L, CHEN J Y, et al. Evaluating modern GPU interconnect: PCIe, NVLink, NV-SLI, NVSwitch and GPUDirect[J]. IEEE Transactions on Parallel and Distributed Systems, 2020, 31(1): 94-110.

[27] ARSID R. Ultra Ethernet and UALink: next-generation interconnects for AI infrastructure[J]. IJSAT-International Journal on Science and Technology, 2025, 16(2): 3103.

[28] LU Y F, GU H X. Flexible and scalable optical interconnects for data centers: trends and challenges[J]. IEEE Communica-

tions Magazine, 2019, 57(10): 27-33.

[29] JIANG L, YAN L S, YI A L, et al. Integrated components and solutions for high-speed short-reach data transmission[J]. Photonics, 2021, 8(3): 77.

[30] NIKDAST M. Silicon photonics for high-performance computing: opportunities and challenges![C]//Proceedings of the 2018 Ninth International Green and Sustainable Computing Conference (IGSC). Piscataway: IEEE Press, 2018: 1.

[31] SHAHRIARI N. 1.1 AI era innovation matrix[C]//2025 IEEE International Solid-State Circuits Conference (ISSCC). San Francisco, CA, USA. Piscataway: IEEE Press, 2025: 10-15.

[32] LAGAEV D A, SHELEPIN N A, KLYUCHNIKOV A S. FD-SOI technology: comparison with FinFET and TCAD simulation[C]//Proceedings of the 2021 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (ElConRus). Piscataway: IEEE Press, 2021: 1996-2001.

[33] LI D L. Virtualization and energy management optimization of high speed computer network data centers based on optical switching and network technology[J]. Thermal Science and Engineering Progress, 2024, 55: 102918.

[34] ISONO H. Latest standardization trend and future prospects for 800 G/1.6 T optical transceivers[C]//Proceedings of the Next-Generation Optical Communication: Components, Sub-Systems, and Systems XII. SPIE, 2023: 14.

[35] HE J, LU D L, XUE H Y, et al. Design of a PAM-4 VCSEL-based transceiver front-end for beyond-400 G short-reach optical interconnects[J]. IEEE Transactions on Circuits and Systems I: Regular Papers, 2022, 69(11): 4345-4357.

[36] MOAZENI S. Next-generation co-packaged optics for future disaggregated AI systems[J]. arXiv preprint, 2023, arXiv: 2303.01744.

[37] MANIOTIS P, TERZENIDIS N, SIOKIS A, et al. Application-oriented on-board optical technologies for HPCs[J]. Journal of Lightwave Technology, 2017, 35(15): 3197-3213.

[38] WILLNER A E. Optical fiber telecommunications VII[M]. London: Academic Press, 2020.

[39] PEI Z X, SONG T, WU C, et al. Cross-timestep fault prediction with imbalanced data for optical modules in Internet data centers[C]//Proceedings of the 2024 27th International Conference on Computer Supported Cooperative Work in Design (CSCWD). Piscataway: IEEE Press, 2024: 1789-1794.

[40] PRABU R T, PANDIAN M M, DHANDAPANI A, et al. High-speed integrated optical transceivers for ultra-high modulation data rates in different optical communication applications[J]. Journal of Optical Communications, 2025.

[41] SAXENA N, ROY A, KIM H. Traffic-aware cloud RAN: a key for green 5G networks[J]. IEEE Journal on Selected Areas in Communications, 2016, 34(4): 1010-1021.

[42] SHANLEY T. InfiniBand network architecture[M]. Addison-Wesley Professional, 2003.

[43] HOUGHTON T, KATZ R. High performance mass storage and parallel I/O[M]. New York: Wiley-IEEE Press, 2001.

[44] SUN Z Z, GUO Z C, MA J D, et al. A high-performance FPGA-based RoCE v2 RDMA packet parser and generator[J]. Electronics, 2024, 13(20): 4107.

[45] FREY P W. Zero-copy network communication: an applicability study of iWARP beyond micro benchmarks[D]. Zurich, Switzerland: ETH Zurich, 2010.

[46] ZHANG J, YU F R, WANG S, et al. Load balancing in data center networks: a survey[J]. IEEE Communications Surveys & Tutorials, 2018, 20(3): 2324-2352.

[47] HUANG S, DONG D Z, BAI W. Congestion control in high-speed lossless data center networks: a survey[J]. Future Generation Computer Systems, 2018, 89: 360-374.

[48] XIA W F, ZHAO P, WEN Y G, et al. A survey on data center networking (DCN): infrastructure and operations[J]. IEEE Communications Surveys & Tutorials, 2017, 19(1): 640-656.

[49] CAI Y P, YAN Y, ZHANG Z H, et al. Survey on converged data center networks with DCB and FCoE: standards and protocols[J]. IEEE Network, 2013, 27(4): 27-32.

[50] IEEE. IEEE 802.1Qbb—IEEE standard for local and metropolitan area networks—media access control (MAC) bridges and virtual bridged local area networks—amendment 17: priority-based flow control[S]. 2011.

[51] Cisco Systems. Priority flow control: build reliable layer 2 infrastructure[R].2010.

[52] RAMAKRISHNAN K, FLOYD S, BLACK D. The addition of explicit congestion notification (ECN) to IP: RFC 3168[S]. 2001.

[53] REINEMO S A, SKEIE T, SODRING T, et al. An overview of QoS capabilities in infiniband, advanced switching interconnect, and Ethernet[J]. IEEE Communications Magazine, 2006, 44(7): 32-38.

[54] SCHARF M, KIESEL S. NXG03-5: head-of-line blocking in TCP and SCTP: analysis and measurements[C]//Proceedings of the IEEE Globecom 2006. Piscataway: IEEE Press, 2006: 1-5.

[55] CRUPNICOFF D, DAS S, ZAHAVI E. Deploying quality of service and congestion control in infiniband-based data center networks: Rev 1.00[R]. USA: Mellanox Technologies Inc, 2005

[56] GEBARA N, GHOBADI M, COSTA P. In-network aggregation

for shared machine learning clusters[J]. Proceedings of Machine Learning and Systems, 2021, 3: 829-844.

[57] LAO C L, LE Y, MAHAJAN K, et al. ATP: in-network aggregation for multi-tenant learning[C]//18th USENIX Symposium on Networked Systems Design and Implementation (NSDI 21), 2021: 741-761.

[58] SAPIO A, CANINI M, HO C Y, et al. Scaling distributed machine learning with in-network aggregation[C]//18th USENIX Symposium on Networked Systems Design and Implementation (NSDI 21), 2021: 785-808.

[59] WANG R Q, DONG D Z, LEI F, et al. Roar: a router microarchitecture for in-network allreduce[C]//Proceedings of the 37th International Conference on Supercomputing. New York: ACM Press, 2023: 423-436.

[60] FENG A X, DONG D Z, LEI F, et al. In-network aggregation for data center networks: a survey[J]. Computer Communications, 2023, 198: 63-76.

[61] GRAHAM R L, BUREDDY D, LUI P, et al. Scalable hierarchical aggregation protocol (SHArP): a hardware architecture for efficient data reduction[C]//Proceedings of the 2016 First International Workshop on Communication Optimizations in HPC (COMHPC). Piscataway: IEEE Press, 2016: 1-10.

[62] LIU S, WANG Q L, ZHANG J Y, et al. In-network aggregation with transport transparency for distributed training[C]//Proceedings of the 28th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 3. New York: ACM Press, 2023: 376-391.

[63] ZHU Y B, ERAN H, FIRESTONE D, et al. Congestion control for large-scale RDMA deployments[J]. ACM SIGCOMM Computer Communication Review, 2015, 45(4): 523-536.

[64] MENIKKUMBURA D, TAHERI P, VANINI E, et al. Congestion control for datacenter networks: a control-theoretic approach[J]. IEEE Transactions on Parallel and Distributed Systems, 2023, 34(5): 1682-1696.

[65] KIM H, RYU J, LEE J. TCCL: discovering better communication paths for PCIe GPU clusters[C]//Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 3. New York: ACM Press, 2024: 999-1015.

[66] AHMAD AWAN A, MANIAN K V, CHU C H, et al. Optimized large-message broadcast for deep learning workloads: MPI, MPI+NCCL, or NCCL2?[J]. Parallel Computing, 2019, 85: 141-152.

[67] AWAN A A, HAMIDOUCHE K, VENKATESH A, et al. Efficient large message broadcast using NCCL and CUDA-aware MPI for deep learning[C]//Proceedings of the 23rd European MPI Users' Group Meeting. New York: ACM Press, 2016: 15-22.

[68] AHMAD AWAN A, CHU C H, SUBRAMONI H, et al. Optimized broadcast for deep learning workloads on dense-GPU InfiniBand clusters: MPI or NCCL? [C]//Proceedings of the 25th European MPI Users' Group Meeting. New York: ACM Press, 2018: 1-9.

[69] LEE S, LEE J. Collective communication performance evaluation for distributed deep learning training[J]. Applied Sciences, 2024, 14(12): 5100.

[70] HIDAYETOGLU M, DE GONZALO S G, SLAUGHTER E, et al. Hiccl: a hierarchical collective communication library[J]. arXiv preprint,2024, arXiv:2408.05962.

[71] HIRSCH D P. Internal development and open source strategies as methods for driving technological disruption in the software industry[D]. vienna: Technische Universität Wien, 2025.

[72] ÁVILA OKADA K F, SILVA DE MORAIS A, OLIVEIRA-LOPES L C, et al. A survey on fault detection and diagnosis methods[C]//Proceedings of the 2021 14th IEEE International Conference on Industry Applications (INDUSCON). Piscataway: IEEE Press, 2021: 1422-1429.

[73] ABID A, KHAN M T, IQBAL J. A review on fault detection and diagnosis techniques: basics and beyond[J]. Artificial Intelligence Review, 2021, 54(5): 3639-3664.

[74] CHRYSANTHOU K, ENGLEZAKIS P, PRODROMOU A, et al. An online and real-time fault detection and localization mechanism for network-on-chip architectures[J]. ACM Transactions on Architecture and Code Optimization, 2016, 13(2): 1-26.

[75] KUMARI P, KAUR P. A survey of fault tolerance in cloud computing[J]. Journal of King Saud University - Computer and Information Sciences, 2021, 33(10): 1159-1176.

[76] QIU K, ZHAO J, WANG X, et al. Efficient recovery path computation for fast reroute in large-scale software-defined networks[J]. IEEE Journal on Selected Areas in Communications, 2019, 37(8): 1755-1768.

[77] CHERRARED S, IMADALI S, FABRE E, et al. A survey of fault management in network virtualization environments: challenges and solutions[J]. IEEE Transactions on Network and Service Management, 2019, 16(4): 1537-1551.

[78] HOPS C. analysis of an equal-cost multi-path algorithm: RFC 2992[R].2000

[79] LIAO H, LIU B Y , CHEN X P, et al. UB-Mesh: a hierarchically localized nD-FullMesh datacenter network architecture[J]. arXiv preprint, 2025, arXiv:2503.20377.

[作者简介]

张慧峰（1992– ），女，博士，之江实验室高级工程师，主要研究方向为新型网络体系架构、网络拓扑结构、高性能网络相关技术等。

刘宁春（1994– ），男，博士，之江实验室助理研究员，主要研究方向为新型网络体系架构、故障诊断与健康管理等。

龙卫平（1988– ），男，之江实验室高级工程师，主要研究方向为高性能网络。

陆平静（1984– ），女，博士，国防科技大学计算机学院副研究员，主要研究方向为高性能计算、高性能互联网络、数据中心网络。

邹涛（1974– ），男，博士，之江实验室研究专家、研究员，主要研究方向为新型网络体系架构、网络协议设计优化。

隆克平（1968– ），男，博士，北京科技大学计算机与通信工程学院教授、博士生导师，主要研究方向为新一代网络技术、光互联网关键技术、无线通信技术、人工智能与大数据。

张汝云（1973– ），博士，之江实验室研究专家，主要研究方向为工业互联网和网络通信安全。

朱俊（1981– ），男，博士，之江实验室工程专家、高级工程师，主要研究方向为新型网络体系结构、软件定义网络、网络资源管理、网络协议设计优化等。