专栏：面向大模型的网络技术

面向大模型预训练的智算网络技术研究

王学聪，冀思伟，李聪
（中国电信股份有限公司研究院，北京 102209）

摘　要：随着人工智能的发展，大规模预训练模型在自然语言处理和计算机视觉等领域都取得了显著成果，促进了智算中心的建设。针对面向大模型预训练的智算网络关键技术展开研究，系统梳理了智算网络国内外最新的标准化进展，提出了一种面向智算网络的目标架构，探讨了智算网络关键技术的原理，包括远程直接内存访问（RDMA）、IB（InfiniBand）、基于以太网的RDMA（RoCE）、集合通信等，同时也分析了智算网络目前存在的问题以及未来的发展趋势，在推动智算网络技术发展、指导智算中心建设等方面具有重要意义。

关键词：智算网络；远程直接内存访问；大模型
中图分类号：TP393
文献标志码：A
doi: 10.11959/j.issn.1000−0801.2024167

Research on intelligent computing network technology for large-scale pre-trained models

WANG Xuecong, JI Siwei, LI Cong
Research Institute of China Telecom Co., Ltd., Beijing 102209, China

Abstract: With the development of artificial intelligence, significant achievements are made in various fields such as natural language processing and computer vision through the utilization of large-scale pre-trained models,which promotes the construction of intelligent computing centers. Key technologies related to large-scale pre-trained models in intelligent computing networks were studied. The latest standardization progress of intelligent computing network at home and abroad was systematically reviewed. A target architecture for intelligent computing network was proposed, and the principles of key technologies, including remote direct memory access (RDMA), IB, RoCE, and collective communication, were explored. Moreover, the current issues and future development trends of intelligent computing networks were analyzed. This research holds crucial importance in advancing the development of intelligent computing network technology and providing guidance for the establishment of intelligent computing centers.

Key words: intelligent computing network, RDMA, large-scale model

收稿日期：2024−04−03；**修回日期：**2024−06−14
基金项目：国家重点研发计划项目（No.2023YFB2904100）
Foundation Item: The National Key Research and Development Program of China (No.2023YFB2904100)

0　引言

自 ChatGPT 问世以来，生成式人工智能（artificial intelligence generative content，AIGC）发展突飞猛进，越来越多的科技巨头相继推出了千亿、万亿参数大模型，带动了全球智算中心的发展与建设。超大规模参数的预训练大模型的硬件需求给智算基础设施带来了前所未有的挑战，也对网络基础设施提出了更高的要求[1]。

网络基础设施是智算中心的重要组成部分，通过高性能网络基础设施，可以满足大模型带来的超大规模算力集群、超高网络吞吐、超低网络时延及高可靠性等业务需求，提高智算中心的图形处理器（graphics processing unit，GPU）的算力使用效率、降低训练时间、提升算力训练持续运行时长，进而提升算力基础设施的整体商业竞争力，具有重要意义。

目前智算网络在生态层面和技术层面都存在一些机遇和挑战，本文提出一种智算中心网络的目标架构，以满足智算中心对网络相关基础设施、互联能力、端网协同控制等的需求，并将着重分析面向大模型预训练的智算网络现状、关键技术以及智算网络的未来发展趋势。

1　智算网络通信标准化进展

1.1　国内标准化情况

当前与智算网络相关的标准仍处于起步阶段，中国通信标准化协会（China Communications Standards Association，CCSA）承担了国内主要的智算网络标准化工作，围绕互联互通和基础支撑，系统化布局智算网络总体技术要求、无损协议、广域网无损传输、大模型预训练等方面的标准化研究。

在智算网络总体技术要求方面，互联网与应用标准技术工作委员会/数据中心工作组（TC1/WG4）针对基建基础设施、硬件基础设施和软件基础设施提出智算中心参考架构，用来指导智算中心的建设；此外，面向智算中心建设方、智算网络提供方和智算业务使用方，对智算中心内部网络的功能、性能、开放性、可靠性提出具体要求。网络与业务能力标准技术工作委员会/网络信令协议与设备工作组（TC3/WG2）关注智算中心内部网络协议，详细介绍了当前主流拓扑、不同拓扑下的路由协议、流量调度机制、拥塞管理技术、远程直接内存访问（remote direct memory access，RDMA）传输优化方案等。

在无损协议方面，TC1/WG4 针对基于 RoCE 网络的不同业务面临的挑战，明确各场景下拥塞控制、流量控制、分组转发、路由选择等相关技术要求，并规定了基于 RoCE 的高速以太无损网络设备的测试方法。TC3/WG2 对智算中心端网协同拥塞控制架构和 RoCE 交换机的拥塞控制功能提出要求，以满足智算中心内高吞吐、低时延和零丢包的性能需求。

在广域网无损传输方面，为实现多智算中心协同计算，RDMA 技术进一步演变成广域 RDMA 技术，网络也从原本的数据中心内组网变为跨域组网。针对广域网设备异构、组网复杂、高性能网络实现技术多样的问题，网络与业务能力标准技术工作委员会/新型网络技术工作组（TC3/WG3）规定了广域 RDMA 网络的需求场景、关键方案和技术要求。同时，为了满足广域网无损传输需求，网络与业务能力标准技术工作委员会/网络总体及人工智能应用工作组（TC3/WG1）进一步研究了广域网编排管控技术、传输能力增强技术和物理层、数据链路层、IP 层的拥塞控制技术。

在大模型预训练场景中，不同的大模型使用的参数规模和数据规模不同，组网实现方式也存在差异，因此网络流量的拥塞点有很大差别。单一拥塞控制算法往往针对某些特定的网络环境，无法满足大模型预训练需求。TC1/WG4 制定了

基于大模型预训练的网络拥塞控制技术要求和测试标准，将拥塞控制算法与大模型预训练场景及网络拓扑、协议相结合，为大模型预训练选择合适的拥塞控制技术提供依据。

1.2 国际标准化及产业化情况

国际标准方面，智算网络的标准化工作主要在ITU及IETF开展，国际智算网络标准体系有待进一步完善。2023年，中国联通、中国电信、信通院、紫金山实验室围绕下一代网络演进（next generation network evolution，NGNe）在SG13启动智算立项。ITU-T Y.NCE-DAICC标准研究分布式智算中心在NGNe中的网络增强需求和能力[2]；ITU-T Y.WALNC标准对广域无损网络的控制器提出功能要求，以提高控制器在路径计算、流量调度、流量控制、拥塞控制等方面的能力[3]。

IETF网络工作组积极推动智算网络关键技术研究，已分别针对广域网中无损技术要求和基于RoCEv2的集合通信卸载需求提交2篇个人文稿；2月24日，在IETF 118次会议的在网计算相关产业技术研讨会上，业界专家围绕智算网络关键技术也进行了深入研讨。

为满足AI和高性能计算（high performance computing，HPC）对智能算力日益激增的需求，2023年7月，由Linux基金会牵头，AMD、Arista、博通、思科等公司联合成立了超以太网联盟（ultra ethernet consortium，UEC），致力于从物理层、链路层、传输层、软件层改进以太网技术，在兼容当前以太网生态的前提下，提升以太网的转发性能。此外，谷歌在2023年OCP全球峰会上提出了一种面向未来AI时代的低时延网络传输协议——Falcon，可为RDMA操作提供端到端可靠传输。

2 智算网络业务需求

2.1 智算网络大规模高性能组网需求

智算中心内部网络需要支持数千至数万张GPU卡的互联互通，通常定义千卡以上为大规模，万卡以上为超大规模。智算中心的GPU组网可划分为参数网络、存储网络、业务网络和管理网络。其中参数网络须支持在数千至上万张GPU卡上同时执行并行训练任务，对卡间互联能力有着极高的要求，一般需要独立设计，保证高吞吐、低时延和无损。

由于全局拓扑决定了不同GPU服务器之间通信所需要经过的路径、跳数等因素，直接影响GPU之间的通信路由，因此，参数网络的任意2块GPU卡间的跳数应尽可能少，路由均衡且避免复杂路径，保障网络整体的高性能和可扩展性，支持弹性扩展，保证足够的冗余和容错能力。

2.2 智算网络超高吞吐需求

大规模并行训练效率依赖于GPU卡间的网络吞吐能力，需要高性能集合通信和高性能网络I/O能力的支持，前者通过集合通信软件库、网络拓扑和网络协议的共同配合实现，后者主要通过引入RDMA技术实现[4]。集合通信的目的是在多个GPU卡之间实现数据或梯度的交换和聚合，以保证模型参数的更新和一致性。集合通信会在同一时间产生大量的通信数据量，以千亿参数规模的AI模型为例，服务器内和服务器间的部分集合通信会产生百GB量级的通信数据量，这些数据量需要在尽可能短的时间内传输完成。为避免集合通信引起的带宽争用、拥塞、冲突等问题，还需要在网络拓扑、拥塞控制和负载均衡技术等方面做更多的优化。

2.3 智算网络超低时延需求

网络时延/抖动影响大模型预训练的效率和质量，智算中心GPU卡间的时延可分为静态时延和动态时延2个部分。静态时延包含数据串行时延、设备转发时延和光电传输时延，静态时延由转发芯片的能力和传输距离决定。动态时延包含交换机内部排队时延和丢包重传时延，通常由网络的拥塞和丢包等不确定因素引起，与网络的负载和

可靠性有关。典型数据中心交换机的硬件转发时延（静态时延）通常在 500 ns～10 μs，在模型预训练业务节点端到端通信时延（通常都在几十甚至上百毫秒）中的占比较小，而动态时延可以达到几十毫秒甚至亚秒级[5]。

分布式训练的集合通信要求时延抖动要尽可能降到最低[6]，网络时延抖动会导致集合通信操作中各个GPU的同步性降低，任意2个GPU之间的点对点通信出现了较大的时延抖动，由于木桶效应，集合通信需要多次等待劣化链路的点对点通信，会放大动态时延和抖动带来的影响。

2.4　智算网络高可靠需求

智算网络的可靠性对模型预训练是否按时按期完成，以及训练质量有重要影响。一是大模型预训练任务一旦中断，重启时间很长，而且会导致部分训练结果丢失或损坏，造成资源浪费、时间延误甚至训练失败；二是大规模智算网络容易出现各种网络故障，如链路断开、交换机宕机、网卡故障等，通过高可靠网络技术，可以最大限度减少故障影响范围和时间，保证集群的计算连通性和完整性；三是大模型预训练需要保持网络性能的稳定和平衡，网络故障引起的性能波动会影响训练效率和质量，导致收敛速度变慢或不收敛。

因此，智算中心网络需要支持冗余链路、备份设备、快速路由切换等技术，实现快速的故障发现和恢复，减少故障影响范围和时间[7]。具备性能波动抑制能力，建立拥塞控制和避免机制，合理分配和调节流量，减少网络性能波动的发生和影响，提高网络的吞吐量、平衡性和稳定性，保证集群的计算效率和质量。

3　智算网络目标架构

为满足智算中心内部网络超大规模、超高吞吐、超低时延、超高可靠性的性能需求，提出了一种面向智算中心网络目标架构，如图1所示，该网络架构包括基础设施层、互联能力层、端网控制层和集合通信层。

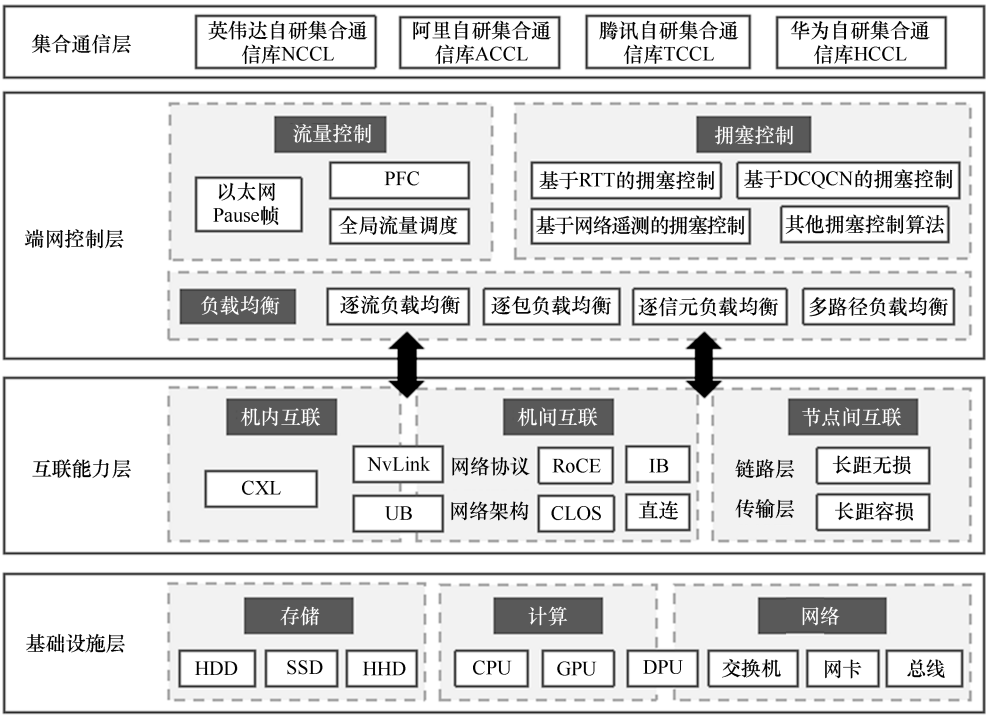

图1　智算中心网络目标架构

基础设施层为承载智算业务的硬件基础设施，包括计算、存储、网络资源等。计算资源主要是指高性能通算和载有GPU或其他加速卡的智算服务器，存储资源包括统一的集中存储或分布式存储等，网络资源则包括机间互联的支持IB（InfiniBand）或RoCE的交换机、网卡和机内总线等。

互联能力层可以分为机内互联、机间互联和数据中心（data center，DC）间互联，机内互联指的是服务器内部的硬件互联拓扑，目前最成熟的代表是NvLink；机间互联指的是一个DC内用于训练大模型的一组设备之间的互联，参数面网络有低时延、高吞吐的要求，传输协议一般采用IB或者RoCE，网络拓扑架构一般为CLOS架构或者直连架构；DC间互联指的是城域内多个数据中心之间的互联，分为长距无损和长距容损，长距无损适用于单集群多DC互联场景，将多个小DC组成一个大的集群，实现算力资源整合，这个距离目前不会太长，而长距容损则适用于多集群多DC之间超长距互联，一般可超过2 000 km。

端网控制层是通过端侧（服务器网卡）和网侧（交换机）配合，采用流量控制、拥塞控制、负载均衡等技术实现端网的高效协同，进而提升智算网络端到端承载的效率。

集合通信层提供大模型预训练的集合通信操作，主要为Send、Receive和AllReduce，目前业界主要的集合通信库除了实现集合通信算子外，还添加了对自家加速卡或网卡等产品的支持，以及定制化场景的算法优化、故障感知等功能。

3.1 基础设施层

基础设施层的计算、存储、网络等资源共同构成了智算的硬件基础设施[8]，计算资源除通算服务器外，还需要包含携带GPU卡的服务器，以完成模型预训练中的大量密集参数运算。存储资源主要用于保存训练中需要的海量样本，以及训练过程中产生的Checkpoints文件等，由于训练过程中会不断读取样本，因此对存储资源的吞吐和时延有一定的要求。网络资源包括服务器之间互联的网卡、光纤和交换机等，智算网络中的网卡均为支持RDMA的网卡，并逐渐向数据处理器（data processing unit，DPU）网卡的方向发展，网卡上集成了拥塞控制算法，与交换机一起协同处理网络拥塞问题。

3.2 互联能力层

机内互联在模型预训练中发挥着重要作用。以NVIDIA为例，服务器内使用了NvLink和NvSwitch实现GPU间流量的高速吞吐。第三代Hopper架构的NvLink速率可达到900 Gbit/s，中间通过NvSwitch互联，互联的架构设计使得机内GPU之间的通信不再依赖PCIe直通Kernel，大大减少了中央处理器（central processing unit，CPU）的负担。机间GPU之间的通信则需要通过PCIe，经过网卡传输数据，NVIDIA服务器内高速互联拓扑如图2所示。

目前智算中心内普遍使用的机间互联网络拓扑是胖树（Fat-Tree）架构。Fat-Tree是一种CLOS架构，分为核心层、汇聚层和接入层。传统的树状网络拓扑中，带宽往往是逐层收敛的，网络带宽并非1:1，而Fat-Tree是无带宽收敛的，每个节点都可以保证上行带宽和下行带宽一致，每个节点都提供了相同的转发能力，因此可以构建大规模无阻塞网络。然而，Fat-Tree作为一种传统CLOS架构，在面向大规模预训练模型场景时也存在一些缺陷，例如，扩展规模受限、大量连线导致的成本问题等。

Dragonfly组网架构是一种直连拓扑架构，因其网络直径小、成本较低的特点，被广泛应用于HPC领域。Dragonfly每个网络节点均有终端节点与之直接相连，没有专门用于网络节点间互连的网络设备。直连架构一般采用自适应智能选路，选择灵活的转发路径，并且可叠加无损技术，提供高性能低时延能力。但目前直连架构主要应用在HPC领域，很少在智算领域有应用案例。

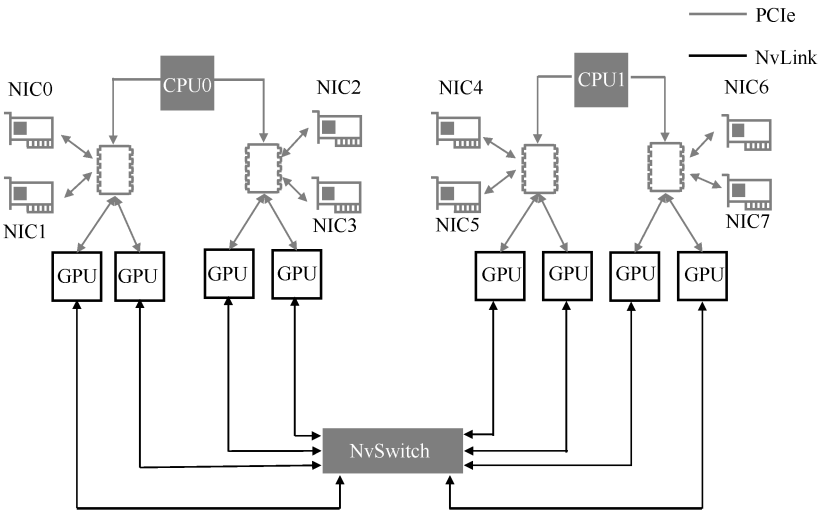

图 2　NVIDIA 服务器内高速互联拓扑

智算中心组网架构优化是一个整体方案，需要综合考虑 GPU 硬件厂商、网络拓扑、集合通信算法和生态等因素。例如 NVIDIA 的 GPU 卡和网络拓扑一般采用多轨道流量聚合技术，实现流量亲和性规划和负载均衡，将通信量较大或较频繁的 GPU 卡安排在相同或相近的物理链路或逻辑通道上，从而减少跨链路或跨通道的通信开销。

单集群多 DC 间的互联可以在单机房服务器容量不足、不能满足模型参数规模时，将分散在一个城域内的多机房互联起来，实现算力资源的整合。单集群多 DC 间长距互联场景如图 3 所示，为 2 个机房，Spine 之间通过长距光纤互联，组成一个大集群，以提供更大规模集群能力。但 DC 间的互联也会带来一些问题，例如，长距会导致传输时延增加、丢包等，对距离有敏感性[9]。

针对 RoCE 网络下的长距 DC 间互联，已在实验室环境中模拟了百千米级的集群内 DC 间协同训练测试，验证并分析城域级智算网络方案可行性。实验室采用 8 台智算服务器搭建智算集群，服务器与 2 台 Spine 交换机和 4 台 Leaf 交换机组成 Fat-Tree 网络，并部署 LLaMA2-13B 模型，构造基于 RoCEv2 的无损网络。

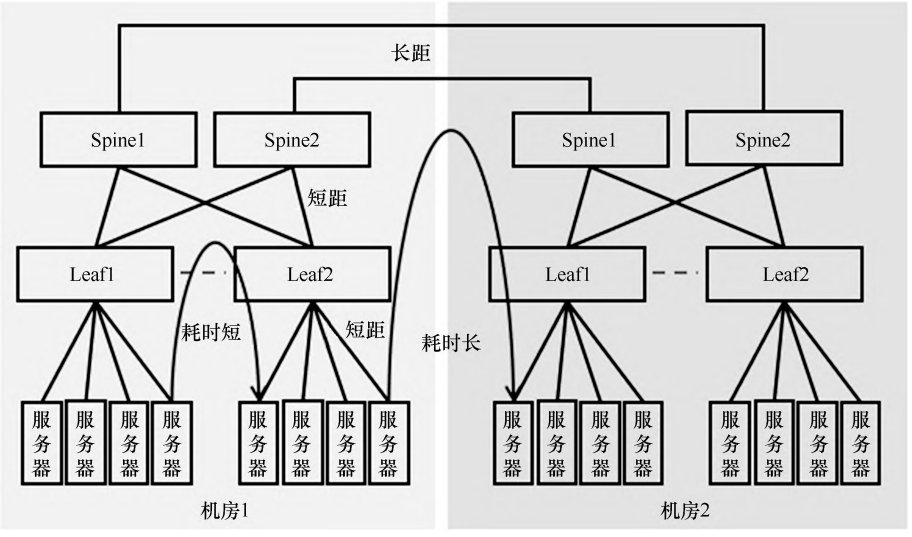

图 3　单集群多 DC 间长距互联场景

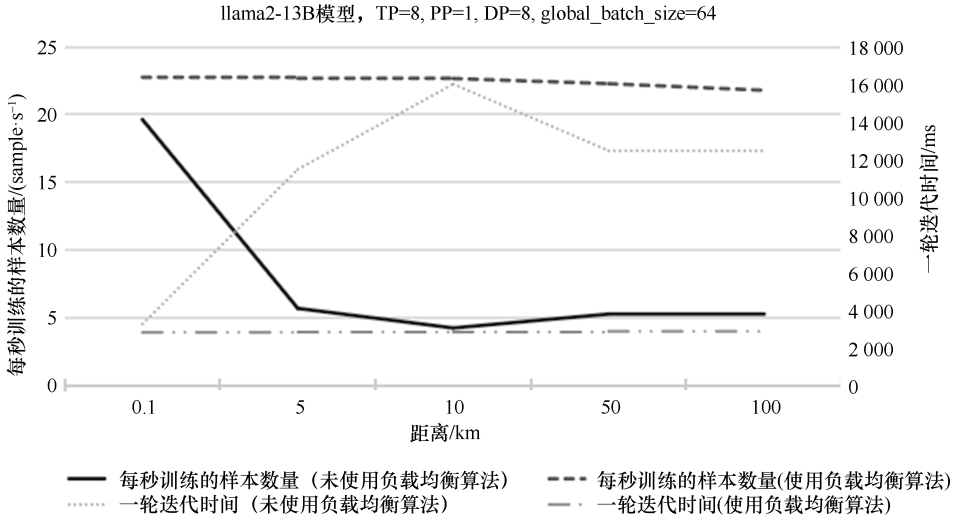

图4　长距DC间互联测试结果

长距DC间互联测试结果如图4所示，在未使用负载均衡技术的网络拥塞场景下，长距相对于短距的预训练性能下降明显，0.1 km下每秒可训练19.63个样本，而5 km下每秒仅训练5.68个样本。由于网络拥塞本身的不确定性，不同距离的性能损失程度呈非线性关系。

在使用负载均衡算法后，随着互联距离的增加，大模型预训练一轮的迭代时间逐渐变长，且每秒训练的样本量呈近似线性下降，预训练性能百千米损失为4%左右。与未使用负载均衡算法的情况相比，长距下利用负载均衡算法可以增加大模型每秒训练的样本数量，缩短模型预训练迭代时间，显著提高GPU利用率。

3.3　端网控制层

端网控制层通过流量控制、拥塞控制、负载均衡等技术维度，解决智算网络端到端承载的效率问题。

流量控制的目的是通过网络控制和全局调度，吸收智算业务中产生的突发流量，主要手段为基于优先级的流量控制（priority-based flow control，PFC）和全局流量调度。其中，全局流量调度通过智能的流量调度机制，将突发流量分散到不同的路径和链路上，避免某些路径过载，从而减轻网络拥塞的风险。全局调度可以根据网络拓扑、带宽利用率和流量负载等因素，动态地分配流量，确保网络资源的均衡利用和高效传输。

拥塞控制主要是通过降低发端速率来避免网络拥塞或者对拥塞的发生作出反应，从而实现端到端高效的拥塞管理[10]。基于往返时延（round-trip time，RTT）的拥塞控制是一种基于RDMA的端到端RTT测量和速率控制的数据中心网络拥塞控制算法，通过RTT测量、速率计算和速率控制3个模块，协同实现拥塞控制；基于数据中心量化拥塞通知（data center quantized congestion notification，DCQCN）的拥塞控制是目前在RoCEv2中使用最广泛的拥塞控制算法，可以提供较好的公平性，实现高带宽利用率[11]。基于带内网络遥测（inband network telemetry，INT）的拥塞控制，利用INT获得精确的链路负载信息并精确控制流量，与传统的拥塞控制相比，基于遥测的拥塞控制技术具有更高的精度和更快的响应速度。

负载均衡技术可以分为逐流负载均衡、逐包负载均衡、逐信元负载均衡和多路径负载均衡等，目前在RoCEv2中普遍使用的是逐流负载均

衡，IB 中使用的是逐包负载均衡，逐信元和多路径负载均衡也属于业界比较前沿的方案。

3.4　集合通信层

在大模型预训练中，机内与机间的流量主要是通过集合通信操作产生的[12]，所以集合通信与网络流量调度、拥塞控制、机内机间互联以及模型预训练效率等有着密切的关系。目前主流的大语言模型预训练方式中，有 3 种主要的并行方式：Tensor 并行、Pipeline 并行和数据并行（data parallelism，DP），其中 Tensor 并行和 Pipeline 并行又合称为模型并行。Tensor 并行将模型的层横向切分，Pipeline 为层切分，DP 则为数据并行。3 种并行方式中 Tensor 并行产生的流量最大，这种流量往往通过服务器内部总线交互；Pipeline 并行产生的流量处于模型预训练的前向传播和后向传播阶段，其特点为点对点的持续流量，这类流量往往需要通过交换机网络传输；数据并行产生的流量处于后向传播完成之后的阶段，流量全部经过网络传输，其特点是瞬时突发流量大、带宽需求高，且需要无损传输。

因此，由于使用的基础设施与互联方式有所不同，业界主流的集合通信库都针对各自的定制化场景做了针对性的优化，例如，ACCL 对 All-Reduce 做了分层优化算法，添加了对自研 RoCE 网络的拥塞感知功能，HCCL 做了定制化场景优化、故障感知等。可以看到，当数据中心中出现新的场景时，例如，算力异构场景、长距互联场景，都会导致原有的集合通信模型每个节点之间不再等价，需要进行定制化优化才能满足大模型预训练的性能需求。

4　智算网络通信协议及关键技术

传统的 TCP/IP 软硬件架构及应用存在网络传输和数据处理的时延过大、多次数据复制、复杂的 TCP/IP 协议处理等问题。RDMA 是一种为了解决网络传输中服务器端数据处理时延而产生的技术，通过提供内核旁路机制和内存零拷贝机制，允许应用与网卡之间的直接数据读写，有效降低协议栈时延，提升 CPU 的效率[13]。RDMA 的性能高度依赖底层网络的无损传输。

基于 RDMA 的通信协议目前主要有 3 种实现方案，分别是 IB、RoCE、iWarp（Internet wide area RDMA protocol）。IB 技术成熟度较高，网络性能好，内置流控技术，对无损支持较好，为英伟达独家控制，成本偏高且生态链薄弱；RoCE 技术相对成熟，网络性能接近 IB，成本适中且生态链丰富，但由于 RoCE 是通过对以太网技术增强实现的无损能力，需要对相关参数进行精细化调教；iWarp 性能弱于 IB 和 RoCE，产业链支持程度不高，逐渐退出竞争市场。

4.1　IB

IB 是一种为大规模、易扩展机群设计的网络通信协议。目前 NVIDIA 的 IB 方案是由一系列技术构成的端到端高速互联网络方案，根据集群规模可部署二层/三层组网，组网采用全连接、轨道优化的无阻塞 Fat-Tree 拓扑，从而形成大带宽、低时延的原生无损网络。IB 的主要特性包括 RDMA、同步管理器、自适应路由（adaptive routing，AR）、逐包转发、乱序重组、SHARP 等。

IB 拥有一套自己的私有协议，与以太网类似，IB 同样采用了分层结构，IB 协议栈如图 5 所示。

其中，物理层定义了 3 种链路速度：1×、4×、12×。根据不同的产品，端口速率也不同，目前 HDR 系列的端口速率可以达到 200 Gbit/s，NDR 系列可以达到 400 Gbit/s。链路层以及传输层是 IB 架构的核心。链路层定义了数据报文的格式，以及报文的转发和本地子网内的交换等。网络层为报文封装了一个全局路由报文头（global route header，GRH），并定义了全局标识符（globally identifier，GID）结构，实现路由转发。传输层将报文传送到某个队列对（queue pairs，

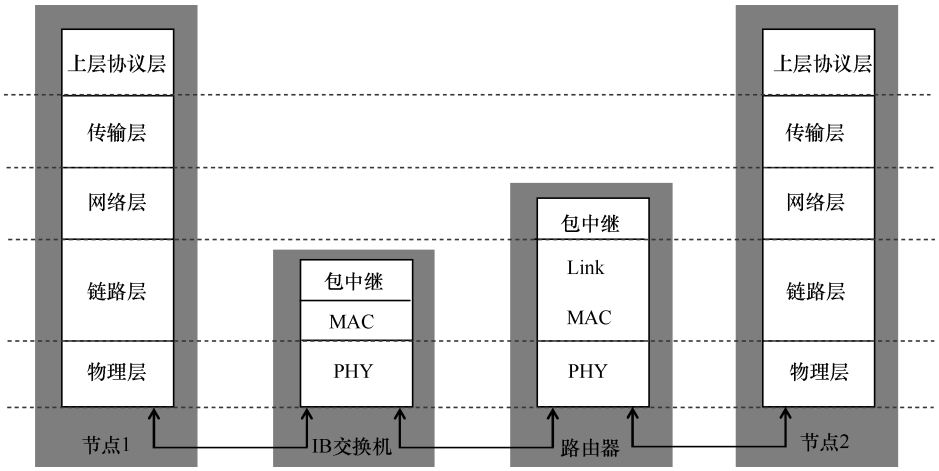

图5　IB协议栈

QP）中，并定义了QP的收发报文操作和RDMA的读写操作等。上层协议层则提供了一些上层协议，例如，SDP、SRP、iSER、RDS、IPoIB和uDAPL等。

IB网络采用基于credit的流控机制（credit-based flow control，CBFC），这种机制只有在确认接收端有足够的credit来接收对应数量的报文后，发送端才会启动报文发送，从根本上避免缓冲区溢出丢包。IB网络中的每一条链路都有一个预置缓冲区，发送端一次性发送数据不会超过接收端可用的预置缓冲区大小，而接收端完成转发后会腾空缓冲区，并且持续向发送端返回当前可用的预置缓冲区大小。依靠这一链路级的流控机制，可以确保发送端不会过量发送报文，网络中不会发生缓冲区溢出丢包。在数据的传输方式上，与以太网的逐流方式不同，IB的数据包传输是以逐包的方式进行，并结合AR和动态负载均衡，可以使流量在交换机上均匀地分布，避免网络拥塞，最大化利用带宽。由于网络中有多条传输路径，这些路径的传输速度和时延可能不同，因此，数据包可能会以不同的顺序到达目的地。为此IB提供了乱序交付（out-of-order delivery）功能，即在IB协议中使用序列号（sequence number，SN）字段来标识数据包的顺序，并在接收端的网卡上根据序列号对数据包进行排序，以确保按照正确的顺序对报文进行重组。

IB的在网计算技术称为SHARP，能够把原先服务器上集合通信中的一部分放在交换机上进行计算，避免了节点之间的多次数据传输，减少了通信数据量，可以有效提升大模型预训练效率。

4.2　RoCE

RoCE在IBTA（InfiniBand trade association）标准中定义，目前共有v1和v2两个版本。RoCEv1和RoCEv2的协议栈如图6所示，RoCEv1基于网络链路层，无法跨网段，应用较少。RoCEv2基于用户数据报协议（user datagram protocol，UDP），可以跨网段，具有良好的扩展性，而且可以做到吞吐、时延相对性能较好，成为被广泛采用的方案。RoCE协议用以太网进行承载，需要辅以PFC[14]、显式拥塞通知（explicit congestion notification，ECN）等技术对传统以太网进行改造，实现低时延和零丢包的无损网络。

PFC在IEEE 802.1Qbb标准中定义，是构建无损以太网的基础技术。PFC将端口流量分为8个优先级队列，每个优先级队列实现独立的Pause机制。当队列已使用的缓存超过PFC门限值时，则向上游发送Pause帧，通知上游设备停

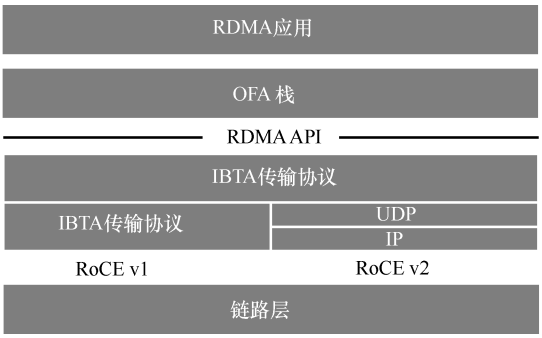

图 6　RoCEv1 和 RoCEv2 的协议栈

止发包，直到上游再次收到 PFC 停止反压报文或经过一定的老化时间后才能恢复流量发送。

PFC 应用示例如图 7 所示，R5 设备收到流量后会给报文分配 cell 资源，当 PFC 功能开启时，会根据报文中的优先级统计占用的 cell 资源。当 cell 资源统计数达到设置的门限值后，再收到该优先级的报文时，会向 R2 和 R5 发送对应优先级的 PFC Pause 帧。R2、R5 收到该优先级的 PFC Pause 帧后，停止发送对应优先级的报文，并对该优先级报文进行缓存。若触发了缓存门限，则再次向上游发送 PFC Pause 帧。

ECN 是 TCP/IP 的扩展，在 RFC 3168 标准协议中定义。当链路发生拥塞时，通过对报文 IP 头中 ECN 域的标识，由目的服务器向源服务器发出降低发送速率的拥塞通知报文（congestion notification packet，CNP），实现端到端的拥塞管理。ECN 是目前以太网中普遍使用的拥塞控制机制，但是 ECN 的水线设置较为复杂。一般 ECN 的水线根据经验设定，并结合现网流量模型进行调整，多次尝试得到合适的阈值。部分交换机能够结合 AI 功能，对现网流量模型进行训练，通过对流量变化的预测，动态调整 ECN 水线值。

4.3　动态负载均衡

RoCE 网络中有效吞吐低的主要原因是大数据流的 Hash 不均。在大模型预训练场景中，网络拓扑要求为 1:1 无收敛，理想情况下，如果流量能在网络中均衡传输，整网可以达到 100% 的吞吐。但是在 RoCE 网络中，传统的处理方式是使用基于报文的五元组进行等价多路径路由（equal-cost multi-path routing，ECMP）的 Hash，很容易出现流量 Hash 不均甚至 Hash 极化，最终导致端口拥堵、链路拥塞、整网利用率低等问题。

针对上述问题，业界在动态负载均衡方面做了诸多尝试，已经有一些私有化实现，但仍处于进一步探索阶段[15]。在流量调度的粒度方面，目前有基于流、子流 Flowlet、包等不同粒度，在封闭系统中甚至还有更小的以 cell 为单位的流量调度。流量调度粒度越小，流量分发越均匀，但小粒度的流量调度容易带来数据包乱序的问题，从

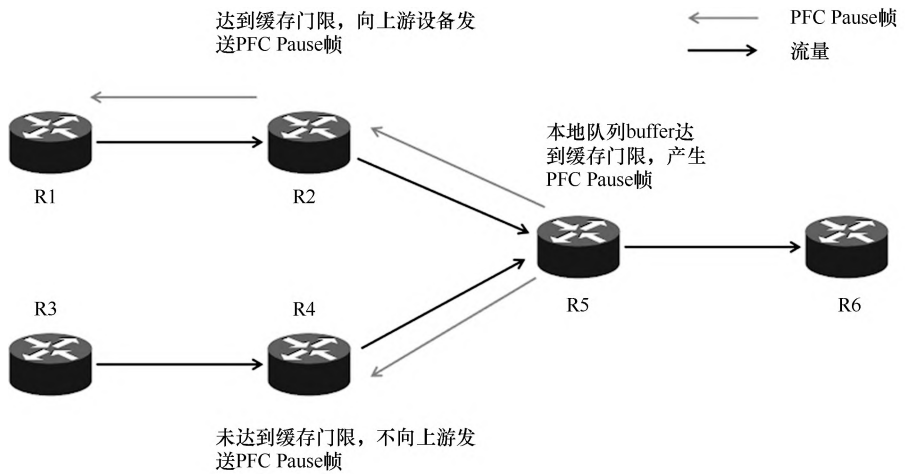

图 7　PFC 应用示例

而对要求保序的传输协议的性能产生影响。逐流Hash均衡是当前最常用的负载均衡算法，基于流量的N元组进行Hash负载均衡，在流数量较多的场景下适用，其优势在于无乱序，劣势在于流数量较少时，会存在Hash冲突问题，网络均衡效果不佳；子流Flowlet均衡是逐流均衡的演进，依赖于子流间时间间隔（GAP）值的正确配置来实现均衡。但由于网络中全局路径及时延信息不确定，因此GAP值无法准确设置，GAP较小存在接收端侧乱序的问题，GAP较大基本等同于逐流均衡；理论上逐包均衡的均衡度最优，但在接收端侧存在大量乱序问题，严重依赖网卡的乱序重排能力。

4.4 集合通信算子

集合通信算子是指模型训练过程中用于各进程之间通信的算子，主要包括Broadcast、Send、Receive、Scatter、Gather、AllGather、Reduce、AllReduce[16]、AllToAll[17]等。集合通信是一组进程之间的多对多通信模型，早期是在HPC领域被广泛使用的信息传递接口。与传统的集合通信相比，深度学习中的集合通信操作依靠硬件加速器和高速互联网络来支持大规模深度学习模型训练，以NVIDIA的NCCL为代表的各类CCL，均沿用了MPI（message passing interface）所定义的通信原语，并在各种场景做了不同的优化。

大模型预训练的3种并行方式中，Pipeline并行的通信是模型层与层之间的数据传递，上一层计算出来的结果作为下一层的输入，使用到的集合通信操作为Send和Receive。Tensor并行和数据并行中使用的集合通信算子均为AllReduce，但是两者处在模型预训练的不同阶段，产生的数据量和对硬件需求也不同。Tensor并行中的AllReduce产生在模型训练的前向传播和后向传播阶段，Tensor并行的目的是把大矩阵运算切割成小的矩阵运算，并均分到各个GPU卡上，再使用AllReduce将计算结果合并。Tensor并行产生的数据通信量在3种并行方式中是最大的，通过机内高速互联网络通信。数据并行的流量产生在模型训练的后向传播完成后，将各个节点计算出来的本地梯度通过AllReduce的方式计算出平均梯度，这个阶段的流量需要通过交换机网络来传输。除此之外，另外一种MOE专家并行模式则使用了AllToAll。

在数据并行阶段，集合通信会产生多对多的瞬时高峰流量，这是产生网络拥塞的主要原因。网络拥塞会降低网络有效吞吐，从而造成训练效率下降，延长训练时间。因此，业内对于AllReduce的算法实现做了很多探索，目前应用较为广泛的是Ring-AllReduce和HD-AllReduce。

5 智算网络未来挑战与趋势

5.1 智算网络面临的瓶颈

智算网络最重要的需求是构建无损网络，而无损网络具有三大核心特征：零丢包、低时延和高吞吐。为此，智算网络需要在流量控制、拥塞控制和负载均衡等方面实现协同优化。IB采取了一系列技术来提升网络转发性能，并且与英伟达GPU自成一体，是智算网络当前的最优选择。但是IB专用的交换机、线缆等设备价格高昂，且生态封闭。RoCE具有更开放的生态和更低的成本，在性能上也接近IB，但是RoCE配置相对复杂，有一定的维护成本，而且在拥塞控制和动态负载均衡等方面还存在很多亟须解决的问题。

在拥塞控制方面，RoCEv2利用PFC和ECN来避免丢包，但是PFC存在PFC风暴、PFC死锁等问题，ECN的水线设置比较复杂。当拥塞出现后，也可能会出现流量反压导致网络有效吞吐变低，从而降低训练效率。

在丢包重传方面，RoCEv2重传机制为回退N帧（go-back-N，GBN），即丢包则从丢包位置之后的N个包都重传。因此，当发生拥塞导致的丢包时，将会进一步加剧拥塞，有效吞吐下降。

在负载均衡方面，RoCEv2 网络仍使用基于流的负载均衡技术。在智算场景中，流的条数少，单流数据量大，因此基于五元组的流 Hash 很容易造成链路拥塞，不能满足智算业务负载均衡需求。

5.2　智算网络发展趋势

无阻塞、低时延、高吞吐成为面向大模型预训练的智算中心内网络的核心诉求。针对智算网络面临的业务需求以及性能、生态上存在的问题，各方积极探索，在拥塞控制、流量控制、负载均衡等方面做了诸多尝试。UEC 致力于打造开放智算生态，在充分借鉴 IB 技术方案的基础之上，从物理层、链路层、传输层和软件层改进现有以太网，同时保留以太网/IP 生态系统的优势，满足了 AI 和 HPC 对智能算力日益激增的需求，得到业界广泛关注。

在拥塞控制方面，UEC 提出现代拥塞控制机制，支持路径拥塞管理、Incast 控制、多场景的灵活适配等。此外，定义了基于 credit 的流控机制，该机制的出现是为了替代 PFC 流控。CBFC 允许接收端将自己的 buffer 空间大小周期性地发送给发送端，从而使发送端可以基于报文优先级和 buffer 大小定量地发送报文，以此来避免拥塞。

在以太网包传输方面，UEC 提出的灵活传送顺序支持 4 种报文传输模式：可靠有序传输、可靠无序传输、可靠用于幂等运算的无序传输和不可靠无序传输，用户可以根据自身需求，选择合适的报文传输方式。

在负载均衡方面，UEC 提出了多路径报文散传的概念，通过实时拥塞管理来选择端到端之间的路径，并提供了细粒度的负载均衡，可实现全路径负载均衡，但是接收端需要有包乱序重排能力。

在网络监测方面，UEC 提出了端到端遥测（E2E INT），基于 INT 的遥测技术相对 RTT 可以更快速地感知拥塞，获取更精确链路负载信息来计算准确的流量更新，从而实现精准控制流量。

6　结束语

本文详细阐述了大模型诞生以来业界对智算网络的需求，提出了一种智算网络目标架构，深入每一项关键技术并详细介绍其原理。此外，对智算网络未来的挑战与趋势做了简单分析，希望能对读者有所启发。在芯片算力强劲发展的今天，智算网络不仅需要在协议、性能、标准等各方面寻求更进一步的发展，还需要兼顾成本、生态等产业链问题，以满足智算中心蓬勃发展的需求。

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]. 2023.
CAICT. China computing power development index white paper[R]. 2023.

[2] ITU. Network capability enhancement for distributed artificial intelligent computing centers in next generation network evolution: TD389[S]. 2023.

[3] ITU. Functional requirements for the controller of wide area lossless network in NGNe: TD294[S]. 2023.

[4] GUO C X, WU H T, DENG Z, et al. RDMA over commodity Ethernet at scale[C]//Proceedings of the Proceedings of the 2016 ACM SIGCOMM Conference. New York: ACM Press, 2016: 202-215.

[5] 百度 . 智算中心网络架构白皮书[R]. 2023.
Baidu. White paper on network architecture of intelligent computing center [R]. 2023.

[6] 熊先奎, 袁进辉, 宋庆春 . 面向分布式 AI 的智能网卡低延迟 Fabric 技术[J]. 中兴通讯技术, 2020, 26(5): 23-28.
XIONG X K, YUAN J H, SONG Q C. Low latency fabric technology of smart NIC for distributed AI[J]. ZTE Technology Journal, 2020, 26(5): 23-28.

[7] 中国信息通信研究院 . 超融合数据中心网络白皮书[R]. 2021.
CAICT. Hyperconverged data center network white paper[R]. 2021.

[8] 中国电信集团有限公司 . 新一代智算数据中心(AIDC)基础设施技术方案白皮书[R]. 2023.
China Telecom. White paper on the new generation of intelligent data center (AIDC) infrastructure technology[R]. 2023.

[9] 赵俊峰, 李芳, 叶晓峰, 等 . 面向广域 RDMA 的确定性网络需

求与技术[J]. 电信科学, 2023, 39(11): 39-51.
ZHAO J F, LI F, YE X F, et al. Research on deterministic networking requirements and technologies for RDMA-WAN[J]. Telecommunications Science, 2023, 39(11): 39-51.

[10] GONG Y Z, ZHANG W, CHEN Y F, et al. How to adapt RDMA congestion control algorithm based on local conditions [C]//Proceedings of the 2023 IEEE International Performance, Computing, and Communications Conference (IPCCC). Piscataway: IEEE Press, 2023: 40-45.

[11] HU Y R, SHI Z, NIE Y, et al. DCQCN advanced (DCQCN-a): combining ECN and RTT for RDMA congestion control[C]// Proceedings of the 2021 IEEE 5th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC). Piscataway: IEEE Press, 2021: 1192-1198.

[12] THAO NGUYEN T, WAHIB M, TAKANO R. Efficient MPI-AllReduce for large-scale deep learning on GPU-clusters[J]. Concurrency and Computation: Practice and Experience, 2021, 33(12): 25-30.

[13] BAI W, ABDEEN S S, AGRAWAL A, et al. Empowering azure storage with RDMA[C]//Proceedings of 20th USENIX Symposium on Networked Systems Design and Implementation (NSDI 23), 2023: 49-67.

[14] CHEN Y Q, TIAN C, DONG J Q, et al. Swing: providing long-range lossless RDMA via PFC-relay[J]. IEEE Transactions on Parallel and Distributed Systems, 2023, 34(1): 63-75.

[15] LUO W F, LAI D H, REN B H, et al. Dynamic load balancing algorithm for distributed database based on PI feedback[C]// Proceedings of the 2022 3rd International Conference on Intelligent Design (ICID). Piscataway: IEEE Press, 2022: 277-280.

[16] LAKHOTIA K, PETRINI F, KANNAN R, et al. Accelerating allreduce with In-network reduction on intel PIUMA[J]. IEEE Micro, 2022, 42(2): 44-52.

[17] CHEN C C, KHORASSANI K S, ANTHONY Q G, et al. Highly efficient alltoall and alltoallv communication algorithms for GPU systems[C]//Proceedings of the 2022 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW). Piscataway: IEEE Press, 2022: 24-33.

[作者简介]

王学聪（1989– ），男，中国电信股份有限公司研究院工程师，主要研究方向为人工智能、云计算、智算网络等。

冀思伟（1996– ），女，中国电信股份有限公司研究院工程师，主要研究方向为云计算、算力网络、智算网络等。

李聪（1993– ），女，中国电信股份有限公司研究院未来网络研究中心副总监，主要研究方向为未来网络技术、下一代互联网技术、数据中心网络等。