面向人工智能的数据通信网络发展

Data Communication Network Development for Artificial Intelligence

高巍/GAO Wei，高静/GAO Jing，杨哲/YANG Zhe
（中国信息通信研究院，中国 北京 100083）
(China Academy of Information and Communications Technology, Beijing 100083, China)

DOI：10.12142/ZTETJ.202406002
网络出版地址：http://kns.cnki.net/kcms/detail/34.1228.tn.20241231.1656.002.html
网络出版日期：2025-01-02
收稿日期：2024-10-28

摘要：基于人工智能技术与业务对数据通信网络的需求，分析现有网络面向数据入算、智算中心互联、大规模AI训练3类场景时存在的问题，阐述“入算”“算内”“算间”网络关键技术创新情况，包括入算网络的业务创新探索，算内网络围绕架构以太网技术等多方面的革新，以及算间网络从IT、IP、光层开展的技术改进，并提出包含运营层、网络管控层、业务连接层、物理网络层的4层网络架构以优化数据通信网络。认为合理推动产业发展需有序规划标准化研究工作，递进式开展关键技术试点验证。

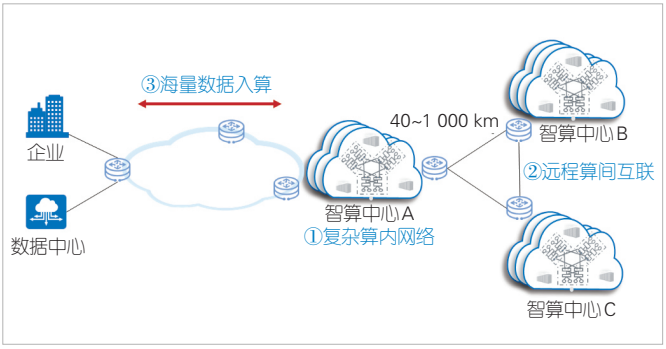
关键词：人工智能；数据通信网络；入算网络；算间网络；算内网络

Abstract: Based on the requirements of artificial intelligence (AI) technology and business on data communication networks, this paper first analyzes the problems of the existing networks in three scenarios: data access for computing, interconnection of intelligent computing centers, and large-scale AI training. The key technology innovation of "access-artificial intelligence data center", "inter-artificial intelligence data center" and "intra-artificial intelligence data center" networks are then illustrated, including the business innovation exploration of "access-artificial intelligence data center" networks, the architecture innovation of "inter-artificial intelligence data center" networks around the Ethernet technology, and the technical improvement of "intra-artificial intelligence data center" networks from IT, IP, and the optical layer. After that, a 4-layer network architecture is proposed, including the operation layer, network control layer, service connection layer, and physical network layer, to optimize the data communication network. We believe that reasonable promotion of industrial development requires orderly planning of standardization research and progressive pilot verification of key technologies.

Keywords: artificial intelligence; data communication network; access-artificial intelligence data center network; inter-artificial intelligence data center network; intra-artificial intelligence data center network

引用格式：高巍, 高静, 杨哲. 面向人工智能的数据通信网络发展 [J]. 中兴通讯技术, 2024, 30(6): 3-9. DOI: 10.12142/ZTETJ.202406002
Citation: GAO W, GAO J, YANG Z. Data communication network development for artificial intelligence [J]. ZTE technology journal, 2024, 30(6): 3-9. DOI: 10.12142/ZTETJ.202406002

数据通信网络作为计算机互联网与电信网的基础网络，历经几十年的发展演进，已从最初依附于电话网络、提供低速而单一的数据通信业务，发展到全球互联的、能够提供数据/语音/视频在内的多种业务的高速宽带网络。近年人工智能、大数据技术的迅猛发展将对数据通信网络应用场景、网络架构产生前所未有的影响。传统大数据中心向智算/算力中心演进，AI训练催生了新的流量模型，带动了海量数据传输的需求。企业/数据中心到智算中心之间的入算网络需要实现样本数据的弹性传输，智算中心内部的网络需要实现数据在计算、存储和网络节点之间的高效流动，智算中心之间的广域互联网络需要实现跨算力中心分布式协同训练场景的无损传输，如图1所示。数据通信网络需要提供全新的“联算”能力，实现算力、算卡的高效联接，从算力使用场景上需要关注入算、算内、算间3张网络的发展[1]。



▲图1　人工智能发展对网络的需求

1 智算业务的发展给数据通信网络提出全新挑战

随着智算业务的发展，传统数据中心向智算中心演进，

网络面临诸多新的需求和挑战。这在用户数据接入智算中心（入算）网络、智算中心内部（算内）网络以及智算中心间互联（算间）网络均有体现。

1）数据入算规模不断扩大。行业开展AI模型训练需要将大量样本数据从企业侧传送到智算中心。以汽车企业为例，进行智能驾驶训练的路测数据量可达60 TB/d，年数据量达到几十PB级。这些数据在上传至智算中心的过程中，既需要相对较大的链路带宽，也会产生大量的大象流。同时，有一些数据涉及企业的敏感信息，不能在数据中心落盘处理，需要随用随传。这给网络的弹性策略、负载均衡策略、安全策略都带来很大的挑战。

2）算内网络规模快速提升。随着AI训练过程中图形处理器（GPU）卡数量的增加，智算中心内部AI集训网络的拓扑愈加复杂。以英伟达最新的GB200计算托盘节点为例，每台服务器的以太网出口带宽将达到200～800 Gbit/s，智算中心内将需要更高速、更复杂、更高质量的网络连接。

3）算间互联需求成为现实。AI大模型参数规模快速扩张，目前已经跨越万亿门槛。据推测，GPT-4的参数量约1.8万亿[1]，预计后续版本参数规模将突破10万亿。相应地，在AI训练过程中投入的GPU加速卡数量也将随之增加。GPT-4使用了2.5万张A100 GPU[1]，而GPT-5的加速卡数量将可能突破10万张。仅从能耗角度而言，单个智算中心已难以承载，多个智算中心之间高速、无损互联成为互联网龙头企业开展大模型训练的潜在需求。

与此同时，网络对智算业务的重要性也显著提升。网络是算力传输的大“动脉”，对智算中心资源利用、集群规模和集群稳定性有着重要影响。首先，智算中心的GPU集群性能与GPU性能并非集群关系，而是跟网络通信的类型和性能有关。网络性能直接决定了算力资源的利用效率。其次，网络设备能力对GPU集群组网规模有一定影响，以典型的Spine-Leaf架构的网络为例，网络设备内部转发芯片容量提升2倍，组网规模便可提高4倍。此外，网络可靠性对GPU集群的稳定性影响重大。GPU集群网络中2%的丢包就会使RDMA吞吐率下降为0[2]。

2 “入算”网络发展

2.1 入算网络业务场景与需求

入算网络连接大量企业、科研机构与算力中心，可快速、高效传递AI训练所需的样本数据，是算力接入的重要“管道”。根据样本数据的敏感程度，智算中心在样本存储和训练过程中可能采取两种不同的策略——非敏感数据落盘处理、存完再训，而敏感数据则可能不落盘、随训随传。因此，入算网络的典型业务场景可分为以下两种：

1） 海量样本数据入算场景

当用户的样本数据为非敏感数据时，数据从本地数据中心上传至智算中心后直接落盘存储，入算网络完成数据样本集从数据源到智算中心计算节点的一次性传输。大模型数据集通常拥有数十GB到PB级的数据量。调研结果显示，交通行业和医疗行业普遍存在PB级数据的上传需求，部分科研机构数据上传周期不定，每次数据量约10 TB到上百TB。部分企业每天上传一次数据，每次数据量达100 TB以上。因此，该场景的入算业务特征主要体现为数据量的庞大。

2） 存算分离拉远训练场景

当用户的样本数据为敏感数据时，数据从本地数据中心上传至智算中心后不落盘存储，入算网络完成数据样本集从数据源到智算中心计算节点的随用随传。政务、医疗、金融等行业涉及公民隐私，大模型训练普遍存在通过网络打通存算、将样本面拉远训练的需求。该场景入算业务特征主要呈现为数据自身的敏感性，以及伴随高频次训练带来的高频次传输与多变的数据量。

2.2 现有网络支持入算业务面临的主要挑战

以上两个场景中，入算业务需要网络在保证传输速度快的同时还要保证传输的安全性。然而，现有网络在应对数据入算业务时还存在诸多问题，主要体现在以下3个方面：

1）接入带宽

当采用传统数据专线进行大数据量样本上传时，百兆专线耗时太长，而万兆专线成本太高。具体如表1所示，针对10 TB、100 TB、1 PB的数据传输量，采用硬盘投递方式的

▼表1 入算数据量时间与价格成本对比

序号	数据传输量	硬盘快递投送方式[3]		100M专线方式		10G专线方式	
		运送时长	价格/元	传输时长	价格/(万元·月$^{-1}$)	传输时长	价格/(万元·月$^{-1}$)
1	10 TB	异地3 d及以上	3 000左右	233 h(9.7 d)	1~2	2.33 h	10~100
2	100 TB	异地3 d及以上	7 000左右	23.3 h(97 d)	10~20	23.3 h	10~100
3	1 PB	异地3 d及以上	73 000左右	23 859 h(994 d)	34~68	238.59 h(9.94 d)	10~100

时长异地3 d及以上；百兆专线传输方式时长分别为9.7 d、97 d、994 d，价格为1万元/月~数十万元/月；万兆专线方式需要的时长分别为2.33 h、23.3 h、2.49 d，专线包月价格约为几十万元/月~百万元/月。综合时长和价格两方面考虑，专线方式同硬盘快递投送方式相比没有明显的优势。

2）网络利用率

智算业务大数据入算的流量多为大带宽的大象流。由于现有网络采用的负载均衡策略主要是以五元组（源IP地址、源端口、目的IP地址、目的端口和传输层协议）来区分流量的，无法识别流量规模，因此大量的大象流在网络中同时出现，会造成网络负载不均衡，从而导致网络利用率大幅下降、算网资源严重浪费。

3）数据安全

一些企业尤其是涉及政务、医疗、金融等对数据隐私要求极高的企业，在进行人工智能模型训练时，不希望自身敏感数据被异地存储，以防造成可能的数据泄露。这对入算网络提出了更高的要求，即需要确保数据传输过程的安全性。

2.3 入算网络业务创新

入算网络承载海量数据到智算中心的高效传输，需要构筑差异化的调度和调优能力，以实现全网大规模节点的多流并发传输，保证整网带宽的充分利用，并满足不同业务入算的服务等级协议（SLA）要求。

目前，三大运营商面向弹性专线积极开展数据快递业务创新探索，各自开展了一些试点项目。

1）中国电信

上海电信着手打造400GE IP弹性无损智算广域网络，通过400GE大容量承载、远程直接内存访问（RDMA）无损传输，以及任务式弹性调度等智算网络技术，提供入算网络服务，并部署AI客户终端设备（AI-CPE），实现10 Mbit/s~100 Gbit/s IP弹性伸缩专线。四川电信基于自研IP业务网架构推出了“超算快线”业务。

2）中国联通

中国联通构建数据要素高效传输基础设施，发布高速数据网络“联数网”。相关场景包括东数西算场景中的海量数据快速传输、区块链和隐私计算中的小量高频数据传输，以及智能网联和AI自动驾驶中的高可靠性和高安全性数据传输等。目前已在部分国企行业数据联数网项目中取得了应用。

3）中国移动

中国移动在2023年发布《中国移动数据快递技术白皮书》[4]，并在2024年7月1日首次上线数据快递业务。该业务依托中国移动算力网络基础设施，结合高吞吐、高可靠、高安全等关键技术，与数据源无缝对接，提供广域、长距、高效的一站式数据传输，适用于大规模数据迁移场景。

三大运营商开展的数据快递业务创新，为入算网络的业务创新奠定了一定的基础，但还有待于结合入算场景的业务特征，进一步构建入算网络能力，拓展业务模式。

3 “算内”网络发展

3.1 算内网络业务场景与需求

算内网络实现智算中心内算卡的互联，可完成单智算中心算卡从百卡到万卡、十万卡的超大规模集群连接，是算力运行的关键“管道”。

算内网络服务于大规模数据处理、人工智能训练和推理等业务场景。为了实现高效的计算和数据传输，算内网络具备大规模组网、高带宽、低延迟、高可靠性和可扩展性。以主要业务场景——生成式人工智能训练为例，其第一性原则就是Scaling law，即大模型的智能水平与模型参数、数据样本和算力3个因素成正比。业界推测，GPT-4参数量约1.8万亿，训练中使用了大约2.15×10^{25} FLOPS算力，训练集群使用约25 000个A100 GPU[1]。随着模型参数量从千亿到万亿、十万亿的增长，模型训练使用的算力卡也从万到十万发展。算内网络只有具备超大规模组网调度、超高吞吐、无损传输、快速故障闭环等能力，才能实现算力效率的100%释放。

3.2 现有网络支持智算训练面临的挑战

传统数据中心网络在支持智算业务对网络规模、带宽、时延、可靠性、运维等需求方面，面临诸多新的挑战。例如：传统数据中心网络的“盒-盒”式组网，无法进行规模升级和演进；流级的负载分担策略，存在有效吞吐低、动态时延大等问题；传统的路由协议，收敛时间过长、端网协同低效；常规的采流和运维技术，对智算业务的大流、高吞吐等特性支撑乏力。因此，为支撑智算业务的快速发展，算内网络关键技术创新势在必行。

3.3 算内网络关键技术创新

算内网络关键技术创新研究主要围绕网络架构、新型以太网技术、高性能集合通信库技术、负载均衡与拥塞控制等技术开展。

1）算内网络架构

算内网络架构的设计和优化是确保算力数据高效传输、

算力资源充分利用的关键。目前研究重点包括：一是高速端口互联，网络系统的光侧采用高速VCSEL、高密度光互连等技术；电侧采用高速电接口、低功耗设计和先进信号处理技术，实现高速、低延迟和高可靠性的数据传输，并可实现400 GE、未来800 GE及以上的高速端口互联。二是网络拓扑优化，组网上增加交换机扇出，采用新型网络拓扑，支撑更加扁平化的组网架构，降低组网成本，提升网络可靠性，以优化数据传输效率。

2）新型以太网技术

算内网络需要处理大量数据并辅助执行复杂的计算任务。RDMA作为高性能网络通信技术，能显著降低传输时延，提升传输效率。由于RDMA对于网络丢包异常敏感，丢包会导致网络性能急剧下降，因此面向算力网络的新型以太网技术主要面向RDMA的技术需求：一是网络拥塞控制算法，通过拥塞控制机制来避免数据包丢失和重传，典型的是数据中心量化拥塞通知（DCQCN）算法，提供较好的公平性，提升带宽利用率和网络吞吐量。二是无损网络技术，RDMA要求网络环境是无损的，即在数据传输过程中不发生丢包，以保证数据传输的性能。业界的超融合以太网络、全调度以太网（GSE）、超以太网联盟（UEC）均通过扩展以太网技术来提升通信效率。

3）高性能集合通信库技术

集合通信是并行和分布式计算中的关键技术，其性能直接影响了分布式任务的速度，决定了集群中所有GPU能否形成合力加速模型训练。现阶段各厂家都采用自有集合通信库，站在第三方立场来看，应围绕以下两个重点方向开展算内网络的高性能集合通信技术研究：一是各类集合通信操作的流量特征的研究分析，如Reduce、All-Reduce、Reduc - eScatter、Broadcast、All-Gather和All-to-All等操作，以及如何通过组合实现更复杂的操作；二是为集合通信操作的仿真提出可行性分析，通过仿真的流量模型，评价智算中心网络性能，为智算中心网络建设方案提供参考意义。

4）负载均衡技术

负载均衡与拥塞控制是确保智算中心网络性能的关键技术。智算场景的流量特征是流数少、单流带宽大。传统基于5元组逐流哈希（HASH）算法的等价多路径（ECMP）技术在流数少的时候极易出现HASH不均的情况。算内网络的负载均衡技术研究的主要方向包括：一是智能调度算法，如动态主流负载均衡（DLB）、全局负载均衡（GLB）算法，实时根据网络流量和节点负载动态调整流量分配策略。二是异构算网协同调度，针对算内可能存在的多种计算资源和网络拓扑，综合考虑不同计算节点和网络链路的性能差异，实现跨平台、跨网络的高效流量调度。三是流量特征识别与优化，有效识别分布式训练通信过程中的“少流”和“大流”、周期性同步突发等特征，采取流量整形、优先级调度等措施提高网络传输效率和计算执行速度。

目前，产业界主要通过打造产业联盟或构建自有体系等方式，开展算内网络关键技术创新：

（1）国际龙头率先打造UEC技术联盟。2023年7月，Linux基金会成立UEC，发布UEC技术愿景白皮书[5]，目前已成立4个工作组并与开放计算项目（OCP）开展合作。该联盟旨在基于以太网，面向大模型和高性能计算场景，从物理层到软件层对以太协议栈和配套芯片产业进行革新，其创始成员包括AMD、Arista、博通、思科、Eviden、HPE、Intel、Meta和微软等全球行业龙头企业，覆盖全产业生态，核心是将“产品”标准化。

（2）中国企业联合发起全调度以太网技术架构（GSE）推进计划。2023年5月，中国移动率联合10余家国内企业率先发布了GSE白皮书[6]，并于同年8月，携手中国信通院，联合华为、中兴通讯、锐捷、新华三等30余家主流互联网，设备商、芯片商、高校院所联合发起GSE推进计划，推动智算中心网络技术创新、标准完善和产业应用，打造高速无损、开放兼容的新型智算中心网络技术体系，助力AI产业发展。该计划的研究范畴涉及物理层、链路层、网络层、传输层、管理和运维体系。目前已在中国通信标准化协会（CCSA）成功推进多个相关行业标准立项。

（3）全球龙头企业构建自有技术体系。以Intel、英伟达、Google、华为等为代表的行业龙头企业，凭借各自在芯片或网络方面的优势，打造自有技术体系来巩固和提升行业竞争力。以集合通信库技术为例，Intel的oneCCL、英伟达的NCCL、AMD的RCCL、华为的HCCL，在对计算资源（GPU类型）、网络资源（IB、Nvlink、PCIe、以太网等）、通信方法（All-Reduce、All-Gather、All-to-All等）的支持方面差异迥然。以网络互联协议为例，节点间网络互联协议存在IB、RoCE和以太3条路线，卡间网络存在NvLink和CXL 2条技术路线，技术路线中涉及的技术不尽相同，差异和壁垒共存。

全球产业界的活跃创新为算内网络技术发展开辟了多条技术路径。然而，技术路线不统一必然会造成网络设备兼容性问题，增加设备互通及软件适配的难度，同时也给运维和运营带来挑战。在算内网络技术发展前期，应统筹开展中国标准的研制，为技术创新先行先试及技术方案对比选型提供有效参考。

4 “算间”网络发展

4.1 算间网络业务场景与需求

算间网络用于实现多智算中心间的高速互联，通过广域高吞吐、长距无损协同能力，有效提升算卡资源利用率，并通过算间协同机制，突破地域限制，整合异地算力资源。

算间网络的业务场景主要是跨智算中心协同训练。跨智算中心协同训练是一种分布式训练方式，模型训练过程由多个智算中心共同参与。分布在不同地理位置的智算中心通过网络实现数据和计算资源的连接，并通过有效的数据同步和任务调度机制实现数据和计算资源的协同工作。随着算力需求的快速增长，在机房、电力等条件受限的情况下，单体智算中心算力的规模也将受限，通过多智算中心互联来构建多智算中心协同训练能力将成为一个重要选择。跨智算中心协同训练可以实现城市内多智算中心、区域内（如国家算力枢纽区域的不同省份间）和区域间（如国家算力枢纽间）算力的高效协同，整合碎片化算力，提升算卡利用率，从而支撑更大模型的训练，缩短模型训练的时间。

为支持跨智算中心协同训练，算间网络需具备多点组网、跨域调度及广域长距无损等能力。其中，多点组网能力可以将分布在不同地理位置的智算中心连接起来，构建一个高效的计算网络；不同智算中心可能属于不同的管理域（不同的机构、地区或国家），跨域调度能力可以打破管理域界限，对各个智算中心的资源进行统一调配。广域长距无损能力用于确保数据在长距离传输过程中保持完整性，对于高精度的协同训练任务至关重要。

4.2 现有网络支持算间协同训练面临的挑战

算间协同训练对现有网络技术的增强和扩展提出了新的挑战，主要体现在以下方面：

1）长距传输时延对计算效率影响较大

跨智算中心通信时，智算中心间的传输距离在数十公里到上千公里范围，远远超出智算中心内部节点间的距离。对人工智能计算而言，在单轮迭代时间固定的情况下，计算效率的损失与通信的时间成正比，因此跨智算中心传输距离越长，传输时延越长，计算效率损失就越大。

2）长距传输给RDMA通信带来挑战

RDMA设计的初衷是实现低延迟、高带宽的直接内存数据传输。RDMA通信中如果发生丢包，数据完整性就会被破坏，接收端需要等待丢失的数据包重新发送，整体传输延迟就会增加。在长距传输场景下，丢包对RDMA通信的影响更为显著。仿真显示，在100 km以上的长距场景下，RDMA对丢包更为敏感。0.10%的丢包会导致吞吐量下降50%以上。

3）长距拉远易导致拥塞丢包

现有网络支撑长距拉远协同训练存在网络带宽限制、传输时延增加、网络设备性能瓶颈、拥塞控制机制不健全等潜在问题，更容易发生拥塞丢包。特别体现在，智算中心网络如果采用三层FatTree组网架构的话，现有数据中心内二层FatTree组网负载均衡算法将不再适用。传输距离越长，链路状态反馈越慢，现有无损机制就越无法保证长距拥塞不丢包。

4）链路故障的影响增加

智算中心长距互联可能更易引发光缆闪断、插损变大等异常，这对正在进行的大规模数据传输任务（如训练参数的同步、海量数据的备份等）而言，会导致数据丢失或者计算任务出错。此外，网络中单链路、单板的故障也引发长距流量拥塞。这些故障将使跨智算中心训练协同工作受到严重影响。

4.3 算间网络关键技术创新

算间网络是构建高效、稳定、低延迟的算力系统的重要组成部分。我们需要从IT层、IP层和光层3个方面开展研究，以应对长距互联对现有网络技术的挑战。

1）IT技术

（1）异构集合通信算法改进

优化集合通信算法可减少长距链路的流量传输，并能在链路收敛的场景下减少流量拥塞和丢包。优化策略通常从数据聚合与压缩、基于预测的预取与缓存机制等方面来考虑。在长距链路的集合通信场景中，对传输的数据进行聚合有助于减少需要传输的数据量，对聚合后的数据采用数据压缩技术可降低网络传输的负载。利用机器学习等手段对集合通信中的数据需求进行预测，提前将这些数据从数据源预取到距离接收端较近的缓存节点上，可避免部分因链路突发压力导致的拥塞丢包情况。

（2）统一调度技术

统一调度技术用于将不同智算中心分散的计算资源和服务能力整合起来，以支持智算中心复杂的业务流程。该技术具体涉及业务编排、作业调度和故障定位。其中，业务编排需要考虑不同智算中心的资源特点和可用性，实现业务任务与不同智算中心资源的最优匹配；作业调度用于实现子作业到智算中心和计算设备的映射分配，并通过调度策略来提高作业执行效率，通过负载均衡策略确保智算中心间资源利用率的相对均衡；故障定位技术用于确定故障位置和原因，对

业务流程中每个环节的执行情况进行记录和追踪。

2）IP技术

（1）全局负载均衡技术

全局负载均衡技术用于保障智算中心间网络资源的合理利用，减少拥塞丢包现象的发生。通过全局负载分担算法优化，适配跨智算中心长距训练组网场景。根据网络拓扑结构和链路实时状态进行动态路径规划，实时监测各条长距链路的负载、带宽利用率、时延等情况，为数据传输选择最优的路径。此外可通过关键帧识别和流级调度算法优化，实现RDMA关键帧加速，优化拉远训练效率。

（2）精准流控技术

通过流级精准反压，系统可实现单流故障不扩散、流间任务隔离。该技术可对网络中的各个数据流进行精准监测与感知，并在关键节点实时收集这些数据流的相关信息。当监测到某个数据流出现异常情况时，依据预先设定的规则和算法来判定是否触发反压机制。在支持RDMA的算间网络中，精准反压可通过修改相关的流控字段来通知发送端降低发送速率。

（3）可视化运维技术

对智算中心互联链路的流量进行逐包、逐跳、随流的时延和抖动测量，可以清晰呈现智算中心互联链路的实际状态。这样便于运维人员直观了解业务流的网络服务质量，并在测量的同时与业务的SLA指标对比，及时调整网络资源分配、优化链路。

3）光网络技术

光网络是智算中心间互联的物理底座。我们需要开展大带宽传输技术、高可靠故障处理和超敏捷光层管控技术研究，以支持算间大带宽、高可靠的底层互联链路能力。相关主要技术包括800G单波高速、C+L超宽频谱、单纤容量96T、极速倒换、波长交换网络（WSON）自动重路由，以及管控资源池化、电驱光秒拆秒建等。

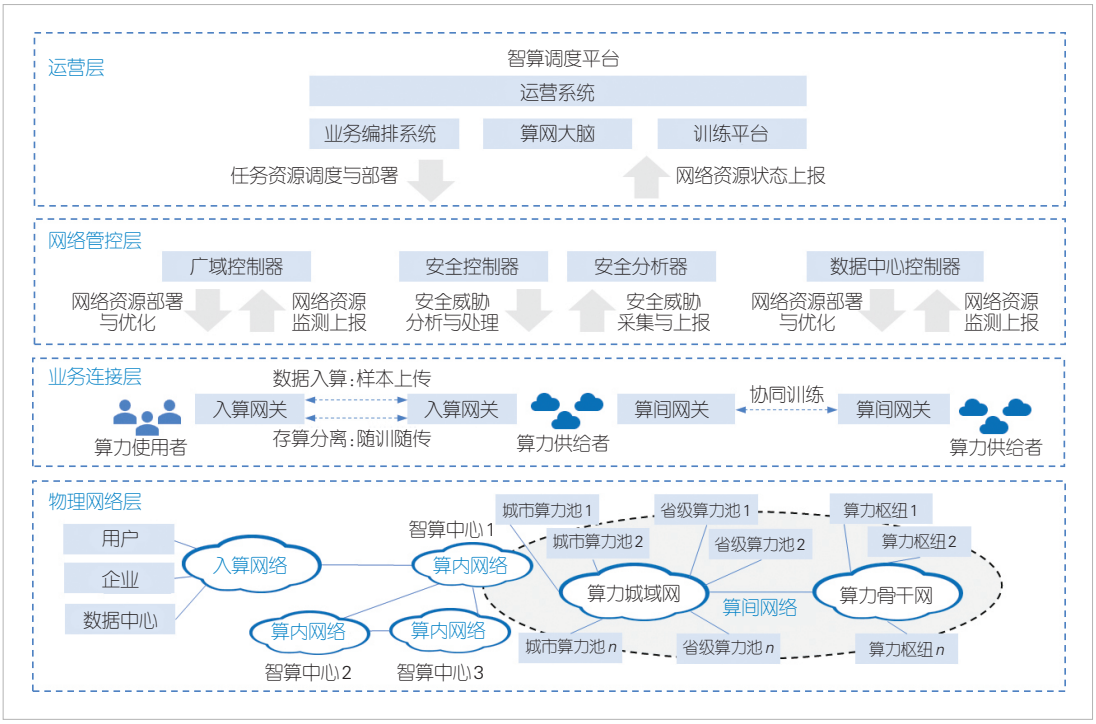
算间网络的发展总体处于早期阶段。目前三大运营商对算间网络的研究工作均有布局，部分互联网企业从自身需求出发开展技术探讨。然而，大模型训练从单智算中心走向多智算中心协同，不仅面临上述技术挑战，还面临建网成本与算效平衡、业务灵活随需调度、多租户安全隔离、故障快速定界定位等商用挑战。

5 面向人工智能的数据通信网络架构与特征

随着人工智能、大数据等技术的飞速发展，算力正逐渐成为数字经济时代的核心驱动力。为助力算力像电力一样成为公共服务，结合入算、算内和算间3部分网络的差异化诉求，综合考虑网络架构、关键技术与创新应用，我们建议采用运营层、网络管控层、业务连接层、物理网络层4层网络架构优化数据通信网络，如图2所示。

1）物理网络层

算内、算间、入算3张网络位于本层。3张网络需求不同，物理位置不同，需要分别使用独立的网络承载。由于现有网络能力无法满足联算的网络需求，推荐新建平面或数据中心单元（POD）来承载联算业务。入算网络连接用户与算力中心，支持2B/2H/2C等用户泛在接入，实现样本数据高品质入算。算内网络支持超大规模组网，具备无损低时延、



▲图2 面向人工智能的数据通信目标网架构

高负载均衡能力，支撑智算集群算力资源高效运行。算间网络实现100~3 000 km多数据中心算力互联，使能多数据中心长距无损协同训练，有效提升算卡资源利用率。

2）业务连接层

对于入算网络，我们需要在企业侧和算力中心部署专门的入算网关。入算网关提供传输层协议转换，为入算流量分配标识并选择合适的隧道和路径，并提供计费对账等能力。此外，入算网关还可为网络的高吞吐传输进行引流，确保流量可以快速入算。对于算间网络，在算力中心部署算间网关，提供RDMA协议联接，为多算力中心协同训练提供超大带宽和长距无损的转发路径。

3）网络管控层

网络资源和安全防护的配置、部署、运维位于这一层，通过网络、安全控制器、分析器，构建网安自治引擎。网络管控层北向对接算力调度运营平台，获取算力任务订阅信息，南向对网络和安全进行规划部署，通过智能引擎分析并计算算力任务所需的最佳网络资源配置和安全防护策略。同时网络管控层能够获取网络/连接层的多层多维信息，构建网络和安全数字孪生，全面提升运维效率。

4）运营层

算力资源的调度、分配、部署，算力服务的业务编排，模型的训练等业务平台位于这个层次。统一的算力调度运营平台让多个业务平台协同服务于算力需求者和供给者。算力调度运营平台南向对接网络管控层，下发任务调度与部署，并获取网络资源信息进行优化调整。

6 总结与展望

面向人工智能的数据通信网络涵盖入算、算内、算间3个部分。其中，入算网络承载海量大数据流入算，需要构筑差异化调度和调优能力，实现全网万级节点的千万流并发，整网带宽充分利用，从而满足不同业务入算的SLA；算内网络需要具备超大规模组网、无损高吞吐以及智能容错能力，实现高算效的释放；算间网络需支持高吞吐、长距无损协同，支持多智算中心协同训练。

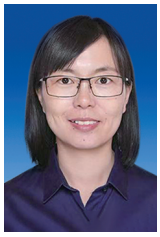
面向人工智能的数据通信网络关键技术创新正成为产业界竞逐的焦点，既包含入算业务模式的创新，又包含节点内、算内、算间网络协议架构方面的关键技术创新，还涉及人工智能技术、“IPv6+”等网络技术与智算中心网络的融合创新。为合理推动产业发展，全面开展网络技术创新，应有序规划标准化研究工作，递进式开展关键技术试点验证。

参考文献

[1] 新质互联网研究组. 新质互联网智鉴报告(V1.0) [R/OL]. [2024-10-12]. https://aimg8.dlssyht.cn/u/551001/ueditor/file/276/551001/1727398847880094.pdf
[2] 中国信息通信研究院. 2024年ICT深度观察 [R]. 2024
[3] 华为云. 数据快递服务DES [EB/OL]. [2024-10-25]. https://www.huaweicloud.com/product/des.html
[4] 中国移动. 中国移动数据快递技术白皮书 [R/OL]. [2024-10-25]. https://www.163.com/dy/article/ICH5G8IN0511BBQE.html
[5] Ultra Ethernet Consortium. Overview of and motivation for the forthcoming ultra ethernet consortium specification [EB/OL]. [2024-10-25]. https://ultraethernet.org/wp-content/uploads/sites/20/2023/10/23.07.12-UEC-1.0-Overview-FINAL-WITH-LOGO.pdf
[6] 中国移动研究院. 全调度以太网技术架构白皮书(2023年) [R/OL]. [2024-10-25]. http://221.179.172.81/images/20230509/69801683620773612.pdf

作 者 简 介

高巍，中国信息通信研究院技术与标准研究所互联网中心主任，高级工程师；主要从事数据通信、云计算、人工智能等方面的研究工作。



高静，中国信息通信研究院技术与标准研究所互联网中心工程师；主要从事数据通信网络方面的研究工作。

杨哲，中国信息通信研究院技术与标准研究所互联网中心工程师；主要从事数据通信网络方面的研究工作。