中文核心期刊　中国科技核心期刊　中国核心学术期刊
第三届国家期刊奖百种重点期刊　信息通信领域产学研合作特色期刊

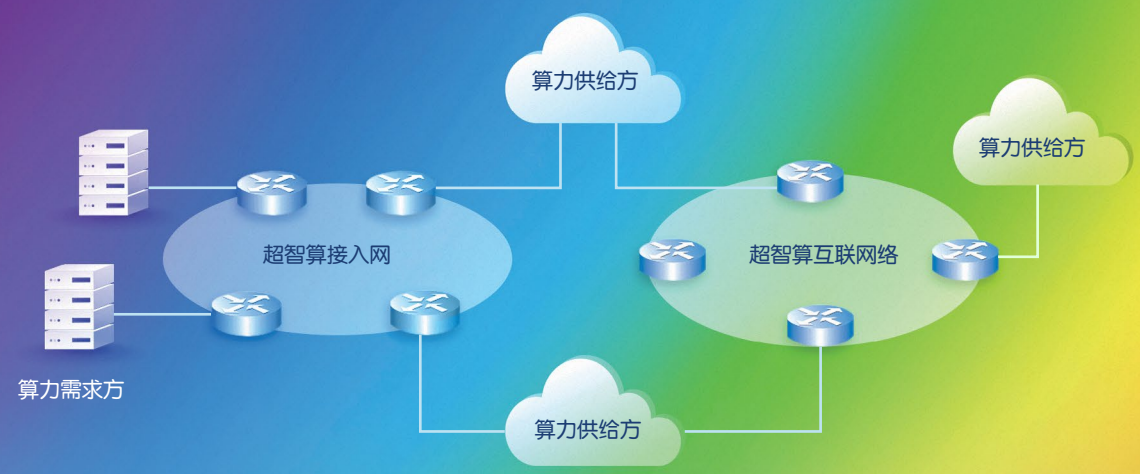
ISSN 1009-6868
CN 34-1228/TN

中兴通讯技术

ZTE TECHNOLOGY JOURNAL

http://tech.zte.com.cn　第 30 卷 · 总第 180 期 · 2024 年 12 月 · 第 6 期（卷终）

专题：数据通信新技术



（封面图片详解见 P11）

ISSN 1009-6868
12>
9 771009 686243

《中兴通讯技术》第 10 届编辑委员会

目　次

中兴通讯技术（ZHONGXING TONGXUN JISHU）
第 30 卷　总第 180 期　2024 年 12 月　第 6 期（卷终）

中文核心期刊 中国科技核心期刊 第三届国家期刊奖百种重点期刊 信息通信领域产学研合作特色期刊 中国知网、万方数据、重庆维普等数据库收录期刊 1995 年创刊

《中兴通讯技术》2024 年热点专题名称及策划人

1. 下一代多址技术
北京交通大学教授 艾渤
北京交通大学教授 陈为

2. 网络大模型
中国电信 IP 领域首席专家 唐宏
中兴通讯无线首席架构师 熊先奎

3. 6G 多天线技术
东南大学首席教授 金石
北京交通大学教授 章嘉懿
东南大学副研究员 韩瑜

4. 6G 无线系统技术
中国信息通信研究院副院长 王志勤
中国移动研究院院长 黄宇红
东南大学教授 王东明

5. 卫星通信技术
哈尔滨工业大学（深圳）教授 张钦宇

6. 数据通信新技术
中国电信研究院教授级高工 解冲锋
中国联通研究院首席科学家 唐雄燕

MAIN CONTENTS

ZTE TECHNOLOGY JOURNAL
Vol. 30 No. 6 (End of Volume) Dec. 2024

期刊基本参数：CN 34-1228/TN*1995*b*16*107*zh*P*¥20.00*6500*15*2024-12

数据通信新技术专题导读

专题策划人

解冲锋

唐雄燕

数据通信是将通信技术和计算机技术相结合而产生的通信方式，是将话音、文字、图像和视频等数据利用网络从一个地方传输到另一个地方的过程。早期存在着包括异步传输模式（ATM）、帧中继（FR）和X.25等在内的多种数据通信技术。此后在各类技术的发展竞争中，以传输控制协议/网际协议（TCP/IP）和以太网为代表的数据通信技术成为主流，特别是基于TCP/IP协议构建的互联网的发展和普及，彻底改变了人类生产生活的方式，互联网也成为人类信息技术设施的核心底座。但近几年来，云网融合、智算组网、确定性网络、安全隔离、天体一体等新型组网形态和应用需求的出现，对互联网的功能和性能提出了更高的要求，也对数据通信技术提出了新的挑战。如何实现更低的延迟、更高的带宽利用率以及更强的安全保障成为了亟待解决的关键技术难题。

本期以“数据通信新技术”为主题，邀请该领域的专家学者撰写了11篇文章，汇集了一系列前沿研究。这些文章介绍并分析了当前数据通信技术的最新关键进展，并对存在的问题和具体的解决方案进行了深入讨论，旨在探索数据通信技术的新突破，并为互联网的未来发展提供有益借鉴。

DOI：10.12142/ZTETJ.202406001
收稿日期：2024-12-01

《面向人工智能的数据通信网络发展》从人工智能技术、业务对数据通信网络的需求出发，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类场景时的问题，总结关键技术发展的趋势与目标网络架构。《高通量数据网演进关键技术》提出了高通量数据网架构及关键技术，通过提高单位带宽下的数据传输体量，解决传统网络传输中遇到的成本高昂和传输时效性差的问题，提升了网络传输通量与效率。同时该文也介绍了在现网开展海量数据超3 000 km传输测试验证的情况。《基于IPv6的虚拟以太网技术——EVN6》提出了一种基于IPv6协议承载的虚拟以太网的新型组网方案。它通过标识映射的方式利用以太网标识和网内主机链路层地址生成IPv6地址，并且选用IPv6地址前缀作为路由信息和站点标识，既标识站点的逻辑位置又使数据包可以通过IPv6的方式穿越互联网。《广域抗损高吞吐URDMA技术》提出一种广域抗损高吞吐超远程直接内存访问（URDMA）技术方案，通过对TCP/IP协议栈的完全卸载，消除CPU对网络高吞吐性能的限制。文章采用拥塞控制、丢包恢复、丢包重传等技术增强标准RoCEv2协议，使其在广域有损网络下保持高吞吐性能。《一种存储高效的IPv6路由查找方法》提出了一种基于前缀拆分模型的集合查找方法（SetSearch），旨在实现高效的IPv6路由查找与存储，并介绍了SetSearch实验评估结果。《智算中心网络技术发展与应用》从应用子层、网卡子层、网络子层以及管控子层构成的完整技术栈出

发，阐述了智算中心网络的关键技术，并在分析智算中心网络发展趋势的基础上，介绍中兴通讯在坚持核心自研的原则下，在芯片、产品和组网方案层面开展的一系列创新研究。《超以太网技术的现状与展望》梳理了超以太网技术的思想、架构和关键技术，探讨了超以太网技术发展面临的机遇与挑战。该技术是面向上述挑战提出的下一代智算以太网技术，旨在从多层次优化传统以太网，提升以太网交换转发性能，改进存储、管理、安全以及遥测能力。《基于生成式人工智能的算力网络自智优化研究综述》讨论了基于生成式人工智能的网络自智优化相关研究进展，提出了生成式算力网络的架构，对其核心流程和所需关键技术进行讨论，并对所提架构的优越性进行仿真验证和分析，最后介绍了生成式算力网络应用场景。《HPN：阿里云大模型训练网络架构》针对大型语言模型（LLM）使得等价多路径（ECMP）极易发生哈希极化，并导致不均匀的流量分布等问题，介绍了阿里云用于LLM训练的数据中心网络架构HPN，并介绍了相关技术。《新型网络芯片技术》针对AI大模型训练等应用对网络互联提出的挑战，论述了包括高性能交换架构、高性能端口、低时延、无损流控、多维负载均衡等新型网络芯片的关键技术，并提出了面向AI的新型网络芯片的发展策略。

“专家论坛”栏目中《网络协议的演进和创新》一文回顾了过去50年主流网络协议的演进过程，分析了传输控制协议/互联网协议（TCP/IP）协议成为当今社会最重要信息设施的根本原因，并指出其协议特征随着网络演进的异化。文章列出过去以10年为单位标志性的网络协议，结合实例讨论协议被大规模采用的原因，并针对加快IPv6部署提出建议。

本期的作者来自知名高校、头部企业与科研机构，面向数据通信新技术，从芯片、网络架构、新型协议、设备的关键技术方面介绍了最新的研究成果。期待这些高质量的研究成果能够为未来网络技术的发展提供有益的参考和启示，并在此对所有作者和审稿专家的大力支持表示由衷的感谢！

策 划 人 简 介

解冲锋，中国电信研究院集团级高级技术专家，教授级高工，中国通信学会会士，中国互联网协会学术委员会副主任委员，北京市IPv6重点实验室主任，曾在美国UCLA大学做政府公派访问学者一年；长期从事宽带网络架构、IPv6下一代互联网、物联网、网络安全、云网融合等方面的研究；参与制定国家IETF RFC标准6项，曾获得2023年度国家科技进步奖一等奖和2023年度中国通信标准化协会科学技术奖一等奖，2019年获得“政府特殊津贴”。

唐雄燕，中国联通研究院副院长、首席科学家，教授级高工，下一代互联网宽带业务应用国家工程研究中心主任，“新世纪百千万人才工程”国家级人选，兼任北京邮电大学教授、博士生导师，工业和信息化部信息通信科技委委员，中国通信学会理事/会士，中国光学工程学会常务理事、会士兼光通信与信息网络专业委员会主任；主要研究领域为宽带通信、光纤传输、互联网/物联网、算力网络、未来网络等。

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。

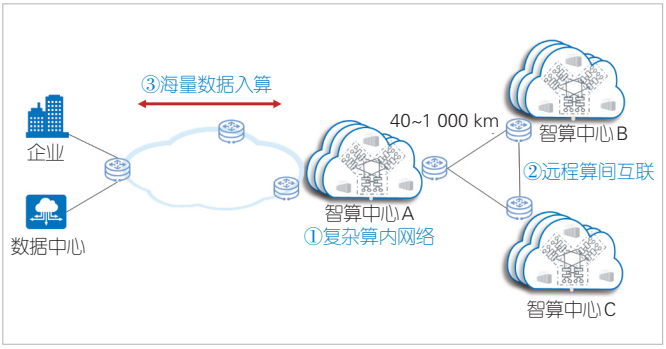
关键词：人工智能；数据通信网络；入算网络；算间网络；算内网络

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。

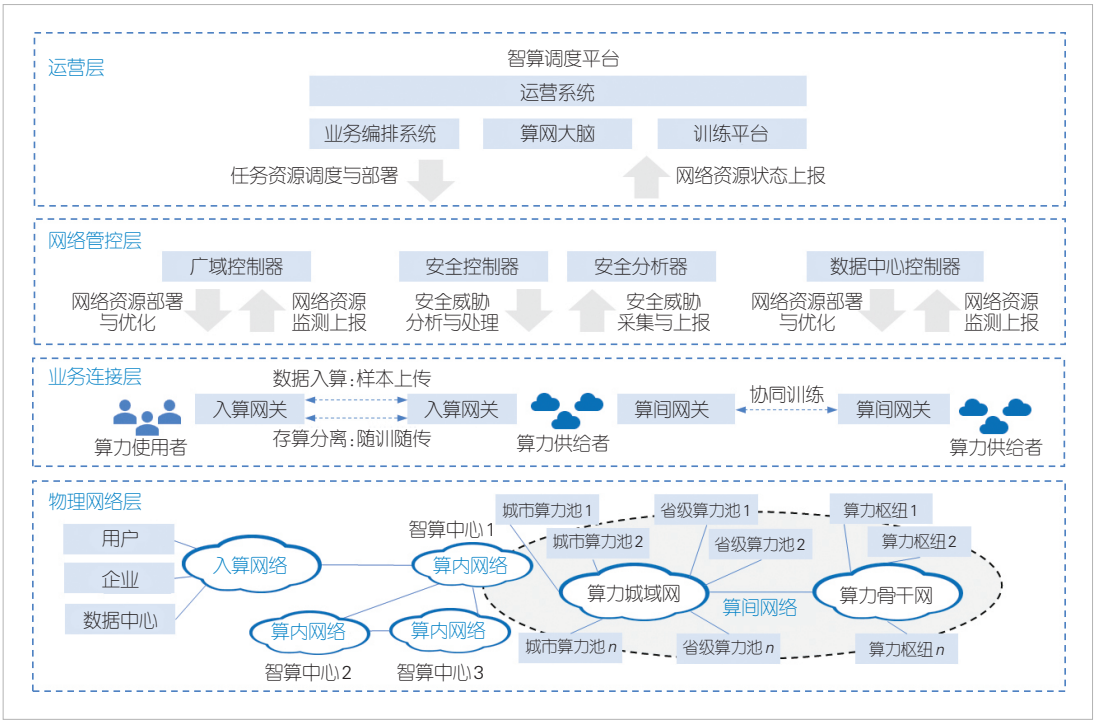
算间网络的发展总体处于早期阶段。目前三大运营商对算间网络的研究工作均有布局，部分互联网企业从自身需求出发开展技术探讨。然而，大模型训练从单智算中心走向多智算中心协同，不仅面临上述技术挑战，还面临建网成本与算效平衡、业务灵活随需调度、多租户安全隔离、故障快速定界定位等商用挑战。

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。

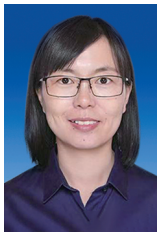
面向人工智能的数据通信网络关键技术创新正成为产业界竞逐的焦点，既包含入算业务模式的创新，又包含节点内、算内、算间网络协议架构方面的关键技术创新，还涉及人工智能技术、“IPv6+”等网络技术与智算中心网络的融合创新。为合理推动产业发展，全面开展网络技术创新，应有序规划标准化研究工作，递进式开展关键技术试点验证。

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。

高通量数据网演进关键技术

Key Technologies of High-Goodput Data Network Evolution

韩梦瑶/HAN Mengyao[1,2]，燕飞/YAN Fei[2]，
曹畅/CAO Chang[1]，庞冉/PANG Ran[1]
（1. 中国联合网络通信有限公司研究院，中国 北京 100048；
2. 中国联合网络通信集团有限公司，中国 北京 100033）
(1. China Unicom Research Institute, Beijing 100048, China;
2. China United Network Communications Group Corporation Limited, Beijing 100033, China)

DOI：10.12142/ZTETJ.202406003
网络出版地址：http://kns.cnki.net/kcms/detail/34.1228.TN.20250109.1329.008.html
网络出版日期：2025-01-09
收稿日期：2024-10-12

摘要：面向智能算力与海量数据复杂交互的高性能传输需求，提出了高通量数据网架构及关键技术，通过提高单位带宽下的数据传输体量，解决传统网络传输中遇到的成本高昂和传输时效性差的问题。采用所提架构中的广域流量调度技术、智能管控运维技术、传输协议优化技术、数据压缩与安全保障等关键技术，实现广域网络弹性带宽分配，提升了网络传输通量与效率。在现网环境开展海量数据超3 000 km传输测试，验证了高通量数据网架构及关键技术的可行性。

关键词：高通量数据网；流量调度；传输协议；弹性带宽

Abstract: In order to meet the high-performance transmission requirements of a complex interaction between intelligent computing power and massive data, a high throughput data network architecture and key technologies are proposed. By increasing the data transmission volume under unit bandwidth, the problem of high cost and poor transmission timeliness encountered in traditional network transmission is solved. The key technologies in the proposed architecture, such as wide area traffic scheduling technology, intelligent management and control operation and maintenance technology, transmission protocol optimization technology, data compression and security assurance, are used to realize the elastic bandwidth allocation of the wide area network and improve the network transmission flux and efficiency. The massive data transmission test over 3 000 km was carried out in the existing network environment, which verified the feasibility of the high-throughput data network architecture and key technologies.

Keywords: high-goodput data network; flow scheduling; transport protocol; elastic bandwidth

引用格式：韩梦瑶, 燕飞, 曹畅, 等. 高通量数据网演进关键技术 [J]. 中兴通讯技术, 2024, 30(6): 10-15. DOI: 10.12142/ZTETJ.202406003
Citation: HAN M Y YAN F, CAO C, et al. Key technologies of high-goodput data network evolution [J]. ZTE technology journal, 2024, 30(6): 10-15. DOI: 10.12142/ZTETJ.202406003

1 算力时代海量数据迁移新需求

近年来，随着大数据、人工智能、移动互联网、云计算等新一代信息技术的应用普及，数字经济发展速度之快、辐射范围之广、影响程度之深，正深刻改变着经济社会的发展进程[1]。算力作为数据处理能力的集中体现，已成为数字经济时代的核心生产力。数字经济的发展带来海量的数据，对数据的存力服务、算力服务和运力服务提出更高的要求。只有数据“存得好”、算力“算得快”、网络“传得稳”，数据才能充分发挥其生产要素价值[2-3]。为了更好地整合算力资源，提高算力基础设施的利用率，算力经济利用云计算、大数据、人工智能等技术，将算力资源集中配置或者部署到云端，从而提高效率和降低成本。因为在此过程中产生了用户侧海量数据的迁移与存储需求，所以如何将用户侧海量的大数据以合理的成本、合理的时效传输到存力/算力基础设施，成为算力经济发展的新需求。随着数据量的增长和数据处理速度的需求增多，越来越多的数据迁移解决方案不断涌现，如东数西算和超智算承载。

1）东数西算

东数西算工程将东部地区的非实时算力需求以及大量生产生活数据输送到西部地区的数据中心进行存储、计算并反馈[4]。按照数据处理对实时性的要求，可将数据分为热数据、冷数据以及介于二者之间的温数据3种。其中，热数据主要包括工业互联网、自动驾驶、远程医疗、灾害预警等产生的需要被计算节点频繁访问的数据，由于热数据对实时性要求较高，所以不适合进行远距离传输；冷数据主要是对后

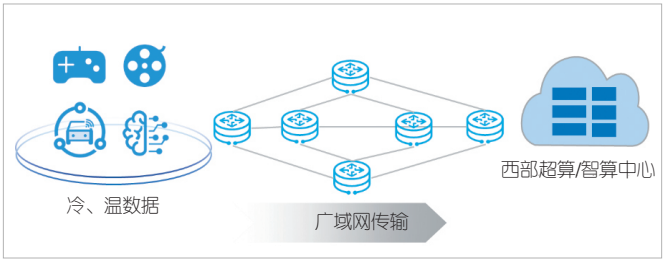
基金项目：中国博士后科学基金资助项目（2024M763570）；国家重点研发计划项目（2023YFB2904201）

台存储、批量备份等存储要求高，但对实时性要求不高的数据。冷数据与温数据非常适合“西算”和“西存”，如图1所示。如何将这些“冷数据”或“温数据”以合理的成本、合理的时效传输到西部存储节点，是目前急需解决的问题[5]。

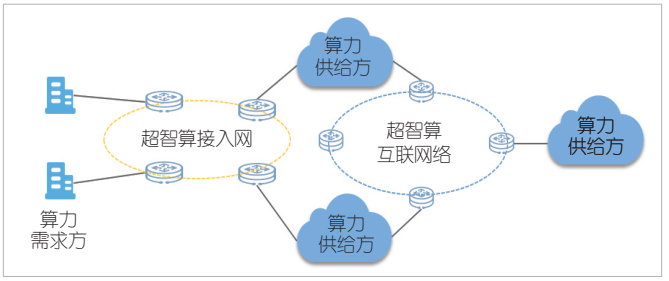
2） 超智算承载

随着网络游戏、影视媒体等业务的快速发展，云游戏、扩展现实（XR）、视频媒体制作等视频渲染需求日益旺盛，需要先通过传输网络数据实时传送到远端算力节点进行演算，再将结果返回到用户侧进行调取使用。此类业务对存力、算力要求高，传输数据量大，如图2所示。为降低算力资源的使用成本，需要将训练数据和训练任务通过网络调度到智算中心进行处理。

海量数据的迁移可以有效整合数据资源和算力资源，但是同时也对传输网络提出新的挑战。目前海量大数据迁移主要有两种方式：通过快递存储介质线下迁移、通过运营商网络线上迁移。但是前者仍然存在着运输成本高、时效性不足、拷入拷出复杂繁琐等问题，同时因为硬盘等存储介质离线搬运，会面临数据损毁、数据泄露等安全风险；后者在应对周期性、临时性大规模数据迁移任务时，存在租用时长无法满足要求、租用大带宽专线成本过高、传输效率受限等问题。不同数据量在不同带宽情况下的理论传输时长如表1所示。



▲图1 东数西算冷温数据传输



▲图2 超算/智算承载场景

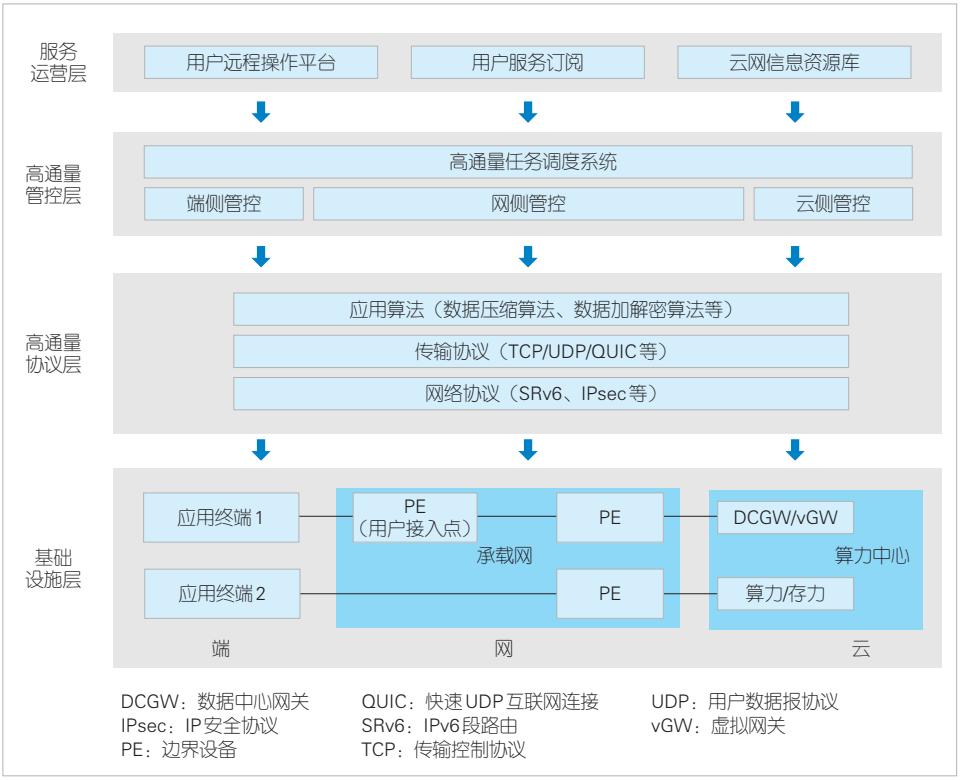
▼表1 不同数据量在不同带宽情况下的理论传输时长

数据量	理论传输时长		
	100 Mbit/s 带宽	1 Gbit/s 带宽	10 Gbit/s 带宽
10 TB	12 d	29 h	3 h
100 TB	121 d	12 d	29 h
1 PB	1 243 d	125 d	12 d

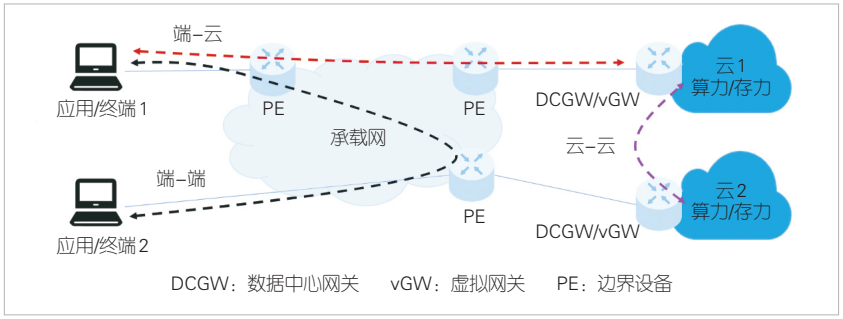
2 高通量数据网架构

高通量数据网通过构建服务运营、高通量管控、高通量协议和基础设施4层架构[6]，提供高通量大数据传输能力，满足算力时代各种应用场景下的数据迁移、同步、协作等需求，实现传输效率与成本的最佳匹配，如图3所示。

基础设施层提供支撑高通量数据传输所需的端侧、网络侧、云侧等软硬件资源，是运力的物理载体，它包括应用终端、承载网以及算力中心，如图4所示。基础设施层在用户和算力、存力间构建起一张随需互联、弹性敏捷的运力网络，支撑算力/存力的灵活调度，针对不同用户对算力、存力的需求，提供并匹配最佳的资源和



▲图3 高通量数据网架构



▲图4　高通量数据网业务流

服务。

高通量协议层包括高通量网络协议、高通量传输协议和高通量应用算法。高通量网络协议需要具备增强网络精准感知能力、流量动态调整能力和基于任务的智能管控调度能力，以实现网络带宽资源充分利用和高吞吐传输；高通量传输协议需要通过多路径并发、精细化控制和智能管控等技术实现数据传输优化，确保高效可靠的数据传输；高通量应用算法关注基于智能化应用算法对数据进行差异化加工，通过存算网协同和高效数据压缩机制最大化降低信息熵，以保证数据归一化和兼容性。

高通量控制层通过编排调度与高通量协议的配合，提供基于任务的智能管控调度能力，实现Overlay层面的业务层选路与Underlay层面的快速开通、整网流量均衡、弹性带宽供给，为用户提供高效、优质的高通量数据传输服务。

服务运营层提供面向最终用户的服务订阅和自助操作能力。使用大数据迁移服务的用户，可以通过运营商提供的服务平台自助订阅线上服务。用户操作平台可以对待传输数据，以及待传输数据的时间计划进行管理配置，查看数据传输进度。云网信息资源库可根据企业分支位置、云池资源找到最匹配的网络路径，也可以根据租户的服务水平协议（SLA）要求，推荐最优路径或最优套餐，实现一体化服务订购。

3 高通量数据网关键技术

3.1 广域流量调度技术

3.1.1 SRv6网络编程技术

SRv6技术结合了源路由优势和Native IPv6简洁易扩展的优点，具备强大的可编程能力和可扩展性，与软件定义网络（SDN）等技术结合，可很好地满足业务快速开通、路径确定性编排、高通量数据传输的需求[7]。

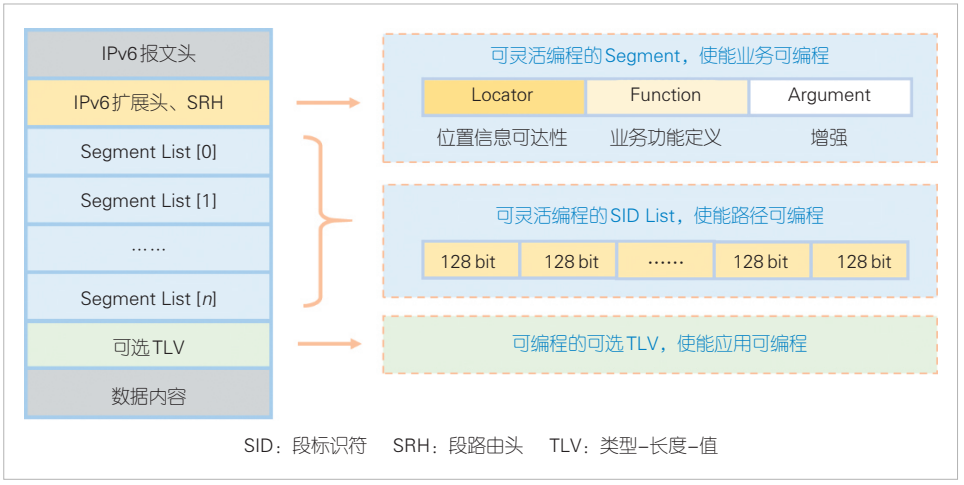
通过结合SDN技术，SRv6可以实现快速灵活的跨域业务开通，简化跨域业务的部署。网络控制器仅需在边界设备（PE）上分别下发SRv6隧道和业务虚拟专用网络（VPN）实例，实现业务的快速配置，使业务开通的时间由几天减少到分钟级，实现用户动态、敏捷、按需的业务开通。

SRv6通过灵活的Segment组合、Segment字段、TLV（指类型、长度和值）组合实现3层编程空间，具备强大的可扩展性[8]，如图5所示。通过源路由机制携带指定转发路径和行为，可实现整网流量路径的统一规划，从而建立满足全局视角的最佳转发路径，最大程度释放网络的价值。

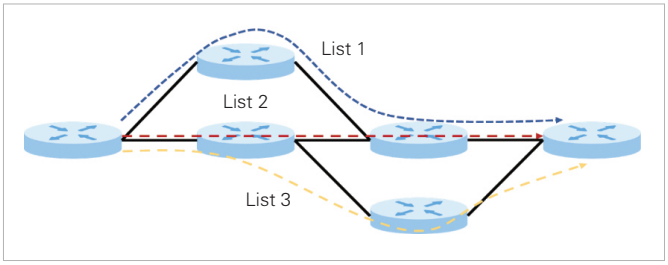
在成本可控的前提下增加带宽资源供给，是实现高通量传输的核心问题。充分利用网络闲置资源，理论上可成倍提升数据传输通量。SRv6的分段路由和源路由特质，使其天然具备流量工程能力。如图6所示，利用SRv6 Policy多List技术，基于整网网络拓扑结构可编排出不重叠的多条路径。基于SRv6的流量工程能力，引导流量从多路径并发传输，从而实现闲置资源的最大化利用。

3.1.2 流量识别与引流

依托当前立体泛在的算力基础设施提供高通量数据迁移服务，需要网络根据不同任务的优先级，提供差异化弹性服务。随着应用感知型IPv6网络（APN6）等技术的发展，可按照不同的网络带宽、时延等SLA需求对任务式高通量数据迁移业务流量进行标识。通过在端侧携带APN扩展头到网络侧，网络侧自动完成业务拆分识别、引流，实现数据迁移



▲图5　SRv6三层编程空间



▲图6 基于SRv6 Policy多List的端到端带宽聚合

业务与普通业务的差异化任务拆分。

通过采用动态网络负载均衡技术，可基于五元组识别出大象流或流组，结合可用资源实时感知技术，分析不同分担路径的实时负载利用情况，进行资源匹配，将高负载路径上的特定大象流调整到低负载路径上。这样不仅可以避免传统哈希算法的缺陷，还能避免路径拥塞导致业务时效性等要求不能满足或利用率低导致带宽浪费的情况，实现全网各路径带宽资源的充分利用，提供精准负载调节能力。

3.1.3 广域拥塞感知与控制

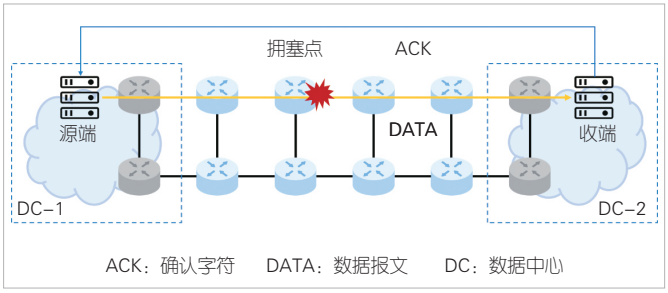
目前网络侧的拥塞状态无法被端侧的传输协议层实时感知，为了保障广域大数据高吞吐传输，需要充分利用网络侧设备的能力，形成新型广域拥塞控制技术，在网络侧直接获取准确的拥塞状态，并根据全局流量信息进行统一速率调控，同时利用反向通告实现亚往返时延（Sub-RTT）控制回路，从而达到及时、准确拥塞控制的效果，如图7所示。

拥塞控制技术的主要实现方式包括网络状态感知和在网流量控制[9]：前者通过利用包括本地队列缓存使用状态和远端RTT变化等信息，使网络设备对拥塞状态进行在网实时感知，作为精细化准确速率控制的基础；后者通过设备感知全局流量状态，并基于此调节反向确认字符（ACK），进行低开销差异化控速，最终实现公平快速收敛。

3.2 智能管控技术

3.2.1 可用带宽资源实时感知

现网中的网络流量呈现明显的时空特征。从空间上看，西部与东部、城区与乡镇、热点区域与稀疏区域之间，在网络流量峰值和均值上都存在较大的差异性；从时间上看，网络资源利用率呈现明显的波峰波谷特征。此外，网络中的不确定性突发也会加剧网络负载的时空不均衡。通过实时感知技术，系统可实时采集网络中各条路径的带宽变化、流量变化趋势、异常事件及可用资源情况，用以支撑网络资源调度。



▲图7 广域传输拥塞感知与控制

3.2.2 流量智能调度

智能管控系统支持根据历史带宽数据、分时采样数据等，预测未来背景流量峰值和持续时间、潜在拥塞点等信息，计算出最佳流量路径，使所计算出来的路径结果能够应对较长时间范围内的背景流量波动，降低路径优化频率，提升传输的稳定性。该系统同时还支持基于用户需求和算力全局信息的最优路径推荐和最优套餐推荐，通过网络多路径编排、网络动态负载均衡、端网协同高吞吐等技术实现高通量数据传输，满足业务多样化体验要求。

3.3 传输协议优化技术

3.3.1 基于TCP的传输协议优化

传输控制协议（TCP）当前的拥塞控制方案影响网络吞吐的主要挑战有：拥塞控制机制相对保守，在高延迟或丢包较多的网络环境中，传输性能可能会受限；一个数据包的丢失或损坏会影响整个数据流的传输速度和效率；拥塞控制回路长，拥塞感知不及时；拥塞感知信息少；端侧拥塞控制动作无差异，端侧局部视角无法感知其他流量状态，无视自身速率，统一速率调整，可能导致不公平、收敛慢。

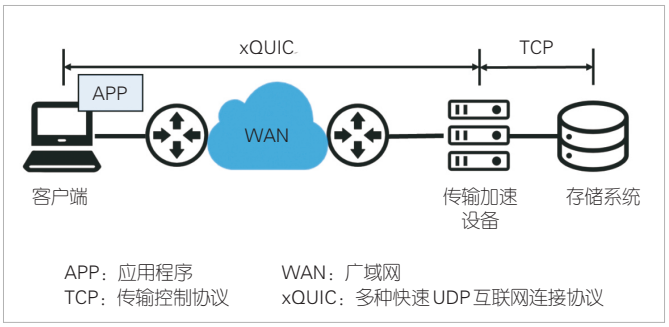
业界基于以上挑战开展了大量基于TCP的传输协议优化研究，例如：通过多路径TCP实现单流多路径数据传输[10]，充分利用网络资源，实现更高速率的数据传输；采用瓶颈带宽和往返传播延迟（BBR）拥塞控制算法，实现数据发送速率控制，即使在轻微丢包的传输链路上也能维持较大的发送窗口，以此提高数据传输的稳定性和效率；通过TCP代理解决终端侧TCP协议栈修改困难的问题，通过分段式传输来提升数据穿越广域网的传输效率。

3.3.2 其他传输协议的优化

业界基于用户数据报协议（UDP）的协议提出了多种优化方案[11]，例如：1）可靠的快速UDP（RBUDP）协议，通

过发送端用TCP发送完成信号表明数据包传送完毕，来实现对所有数据的接收；2）Tsunami协议，通过周期性地对未收到的数据分组发送Negative ACK，同时通过基于丢包率的拥塞控制机制，保证网络传输性能；3）基于UDP的数据传输协议（UDT）协议，设计和实现了功能和效率满足需求的传输协议，同时具备可应用于互联网的拥塞控制算法，保证效率、公平性和稳定性。

以谷歌快速UDP互联网连接（gQUIC）/华为快速UDP互联网连接（hQUIC）等为典型代表的新型传输协议，在文件传输场景具备以下独特优势：快速连接机制、多路复用机制、新的序号和确认机制、纠错机制[12]。快速UDP互联网连接（QUIC）协议可以作为客户端应用程序（APP）与传输优化设备，以及独立传输优化设备之间的通信协议，如图8所示。基于QUIC协议框架和待传输内容的特点，系统可针对性地进行扩展和增强，以满足不同应用场景对传输性能的需求。



▲图8 客户端和云端传输优化设备部署使用xQUIC通信

3.4 数据智能压缩技术

数据压缩可以减少网络内的数据传输量，尤其是在需要穿越广域网络且出口带宽有限的情况下，数据压缩可以显著提升传输速度，缩短传输时间，以实现低传输成本的目标。数据压缩的原理是通过消除或减少数据中的冗余信息来减小数据的表示大小，根据信息恢复完整度可以分为无损压缩和有损压缩两种类型。常见的无损压缩算法有：字典压缩算法，如LZ77、串表压缩算法（LZW）等；熵编码算法，如霍夫曼编码；预测编码算法，如Delta编码。

对于不同的数据传输类型，例如文本、音频和视频，特定的结构化数据可采用不同的策略，也可采用在线学习的方式，根据当前的传输环境和整体的传输要求，动态地选择最适合的压缩算法，以适应不同数据传输类型的需求。在软件定义框架下，压缩算法的定义和编排由管控系统来完成。管控系统根据传输任务的要求和网络环境的实时情况，选择合适的压缩算法，并将其应用于相应的数据传输节点。这种灵活性和自动化的决策过程使得数据传输系统能够根据实际需求和情况进行优化，提升整体数据传输的效率。

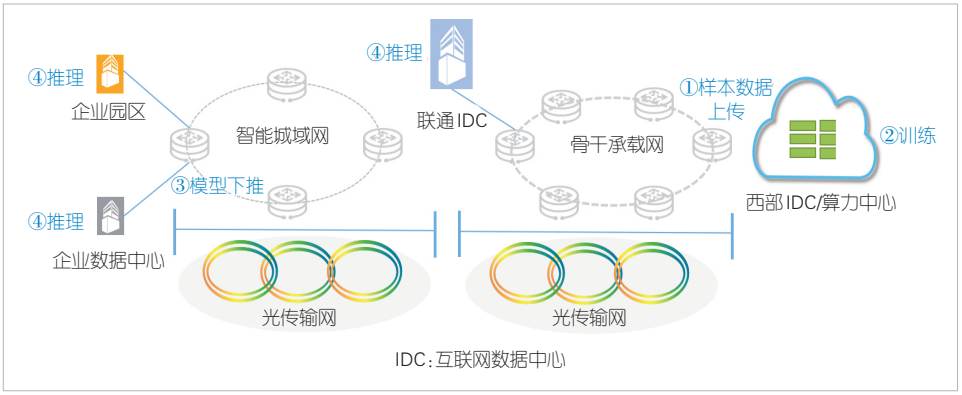
3.5 数据传输安全保障技术

为了保障敏感业务的传输安全，可以通过网络切片技术实现专属资源转发。转发面使用FlexE灵活以太技术实现细粒度带宽资源硬隔离。控制面采用分布式网络切片控制协议标识设备硬件预留资源，实现安全隔离，保障传输全程可信安全。但此类隔离能力成本较高，用户可按需选择。

以强化网络内生安全为目标，防止网络传输数据窃取、终端设备仿冒、路由劫持和分布式拒绝服务（DDoS）攻击等安全风险，承载网络路由器目前从以下几个方面加强安全防护：1）通过管道加密、管理协议强认证和加密算法等措施，保护用户数据和认证数据安全，例如使用基于IP安全协议的通道认证及加密传输等技术；2）路由器支持路由安全措施，在建立路由协议邻居关系时对身份进行认证，支持HMAC-SHA256高强度认证算法，以及通过keychain动态更改密码链，同时在学习和发布路由时进行校验，确保对等体可信，保障路由路径信息不被篡改；3）构筑协议秒级防攻击能力，保证网络对短时强DDoS攻击的防护能力。

4 高通量数据网传输能力测试分析

面向海量数据高效传输需求，我们开展了实际传输距离超过3 000 km的海量数据广域高通量传输验证。现网测试验证结构示意图如图9所示。基于中国联通覆盖中国的169骨干互联网，本次测试是将上海智算业务训练数据导入宁夏中卫智算训练集群的典型“东数西算”场景。



▲图9 高通量数据网现网验证示意图

针对高通量数据网部分测试数据的简要分析如下：

1）网络弹性带宽。测试实验验证了基于IP承载网络的3 000 km海量数据任务式长距传输，通过任务式传输能力调用IP网络闲时带宽，支持用户传输带宽从100 Mbit/s到N×1 Gbit/s的智能弹性调整（10～300倍）。

2）传输协议优化。通过采用传输协议参数优化、数据智能切分/合并、多流并发技术，相同带宽下超大文件传输效率比现有传输能力提升4倍以上。针对数据传输流量与现网业务流量的带宽资源竞争问题，仍需要研究优先级保障的拥塞控制技术，避免影响其他高优先级业务。

3）多地协同传输。本次测试验证了基于SRv6协议的IP骨干网路径可编程能力，实现了上海–广州–宁夏三地协同、4条并发路径的高通量数据传输，为未来多地协同、超高通量传输奠定了技术基础。

4）存运协同。实验分析了不同硬盘配置对海量数据传输性能的影响。结果表明，网络能力应该与存储能力匹配，同时应充分利用网络的带宽。然而，存运协同相关问题仍需要解决，以进一步提升海量数据的传输能力。

5 结束语

高通量数据网是面向算力时代的运力增强需求。本文中我们提出了网络承载、智能管控、端侧优化协同演进方向，通过3 000 km的海量数据广域传输试验验证了高通量数据网架构的合理性。实验结果表明，广域流量调度、智能管控、传输协议优化等关键技术具有可行性，能够实现网络有效带宽最大化、传输效率最大化、网络丢包最小化、现网影响最小化。面向未来，高通量数据网需要以现有网络为基础，以支撑算网产品化创新为目标，充分挖掘网络潜力，提升网络资源利用率，增强网络的传输能力，助力“东数西算”国家战略的实施落地，以高品质智算互联网赋能人工智能产业，加快形成以数字化、网络化、智能化为特征的新质生产力。

参考文献

[1] 汪玉凯. “数据要素×”与“东数西算”: 全国一体化算力网建设的关键 [J]. 人民论坛, 2024(8): 52–56
[2] 2023中国算力大会. 中国存力白皮书(2023年) [R]. 2023
[3] 华为. 数据存力: 高质量数据发展的数字基石 [R]. 2022
[4] 中国智能计算产业联盟. 东数西算下新型算力基础设施发展白皮书 [R]. 2022
[5] 中国联通研究院. 面向“东数西算”的算力网络关键技术白皮书 [R]. 2022
[6] 中国联通研究院. 高通量数据网架构及关键技术白皮书 [R]. 2023
[7] 华为. SRv6 [EB/OL]. (2024–08–12)[2024–10–25]. https://support.huawei.com/enterprise/zh/doc/EDOC1100193023
[8] SHIMATANI S, KASHIWAZAKI H, NOBUKAZU I. SRv6 network debugging support system assigning identifiers to SRH [C]// Proceedings of IEEE 47th Annual Computers, Software, and Applications Conference (COMPSAC). IEEE, 2023: 518–525. DOI: 10.1109/COMPSAC57700.2023.00075
[9] PENG F, LU B C, SONG L, et al. PACC: perception aware congestion control for real–time communication [C]//Proceedings of IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2023: 978–983. DOI: 10.1109/ICME55011.2023.00172
[10] TRAN V H, SADRE R, BONAVENTURE O. Measuring and modeling multipath TCP [M]//Intelligent mechanisms for network configuration and security. Cham: Springer International Publishing, 2015: 66–70. DOI: 10.1007/978–3–319–20034–7_8
[11] CHOUDHARY G K, KANAGARATHINAM M R, NATARAJAN H, et al. Novel MultiPipe QUIC protocols to enhance the wireless network performance [C]//Proceedings of IEEE Wireless Communications and Networking Conference (WCNC). IEEE, 2020: 1–7. DOI: 10.1109/wcnc45663.2020.9120821
[12] 王继昌, 吕高锋, 刘忠沛, 等. QUIC传输机制与应用综述 [J]. 计算机工程, 2023, 49(6): 1–12 DOI: 10.19678/j. issn. 1000–3428.0065493

作 者 简 介

韩梦瑶，中国联合网络通信有限公司研究院博士后；主要研究领域为下一代互联网架构演进与关键技术研究。

燕飞，中国联合网络通信集团有限公司网络运营事业部工程师；主要研究领域为IP网络技术，负责IP骨干网网络运维等工作。

曹畅，中国联合网络通信有限公司研究院下一代互联网研究部总监，正高级工程师；主要研究领域为算力网络、IPv6+网络新技术、未来网络体系架构等研究工作。

庞冉，中国联合网络通信有限公司研究院数据通信首席研究员；主要研究领域为下一代互联网架构演进与关键技术研究。

基于IPv6的虚拟以太网技术——EVN6

IPv6-Based Ethernet Virtual Network (EVN6)

马晨昊/MA Chenhao，孙吉斌/SUN Jibin，
解冲锋/XIE Chongfeng
（中国电信股份有限公司研究院，中国 北京 102209）
(China Telecom Research Institute, Beijing 102209, China)

DOI：10.12142/ZTETJ.202406004
网络出版地址：http://kns.cnki.net/kcms/detail/34.1228.TN.20241217.1409.002.html
网络出版日期：2024-12-17
收稿日期：2024-10-09

摘要：针对传统虚拟以太网组网技术的封装开销大、组网范围有限和数据面能力较弱等问题，提出了一种新的直接基于IPv6协议承载的虚拟以太网组网方案。该方案可以使以太网帧直接封装在IPv6报文的负荷上，通过标识映射的方式利用以太网标识和网内主机链路层地址生成IPv6地址，并且选用IPv6地址前缀作为路由信息和站点标识，既标识站点的逻辑位置，又使数据包通过native IPv6的方式穿越互联网，提升了封装效率和可运维性。现网测试验证了技术的可行性，展现了其在业务敏捷性、覆盖范围和流量转发调度方面的优势。

关键词：IPv6；虚拟以太网；标识映射

Abstract: The traditional virtual ethernet networking technology has the problems of high encapsulation cost, limited network coverage, and weak data plane capability. In this paper, a new virtual ethernet networking system based on IPv6 protocol is proposed, called IPv6-based ethernet virtual network or EVN6. It can make the ethernet frame directly encapsulated on the payload of IPv6 packets, generate IPv6 address using the information of the virtual Ethernet and the host by identification mapping, and select the IPv6 prefix as the routing information and site identification, which not only identifies the logical location of the site, but also enables the packet to traverse the Internet through native IPv6. This system improves encapsulation efficiency and operability. The technical feasibility of technology is verified in the field trail test, showing the advantages of business agility, network coverage, and flow forwarding and scheduling.

Keywords: IPv6; virtual ethernet; identification mapping

引用格式：马晨昊，孙吉斌，解冲锋．基于IPv6的虚拟以太网技术——EVN6 [J]. 中兴通讯技术，2024, 30(6): 16-22. DOI: 10.12142/ZTETJ.202406004

Citation: MA C H, SUN J B, XIE C F. IPv6-based ethernet virtual network (EVN6) [J]. ZTE technology journal, 2024, 30(6): 16-22. DOI: 10.12142/ZTETJ.202406004

随着云计算、虚拟化技术的快速发展，虚拟以太网成为现代网络架构中的重要组成部分，广泛应用于运营商的城域网、骨干网、企业组网、数据中心、家庭网络和软件定义广域网（SD-WAN）等多样化场景。虚拟以太网是一种基于虚拟化技术的网络连接方式，它在虚拟机之间、虚拟机与物理机之间提供了高效的网络通信能力。当前虚拟以太网类技术主要包括基于多协议标签交换（MPLS）的二层虚拟专用网络（VPN）（包括VPWS、VPLS）[1]、基于IP的可扩展虚拟局域网络（VXLAN）[2]、EVPN VPWS/VPLS over SRv6 BE等[3]。

虚拟以太网具有众多优点：1）具有灵活性，提供了灵活的网络配置能力，使用户可以根据业务需求动态调整网络拓扑和流量策略；2）具有良好的可扩展性，通过VXLAN等技术，可以支持大规模网络部署，适应数据中心和云计算平台的扩展需求；3）通过冗余和故障转移机制，提高了网络的可靠性和可用性；4）通过提高资源利用率和简化网络管理，降低了企业的IT成本。

当前的虚拟以太网满足了二层虚拟组网的基本要求，如在云数据中心中支持虚拟机迁移，支持多种虚拟网络拓扑模式，包括点到点、多点到多点等。在虚拟网之间的安全隔离，确保租户之间数据和拓扑不可见、相互不干扰，可为不同的租户或者网络内业务提供差异化的服务质量保障。

随着IP协议第6版（IPv6）的大规模部署，网络已基本进入双协议栈运行时代[4]。IPv6网络的覆盖范围已经延伸至

5G[5]、固定宽带、数据中心和云等网络业务。端到端打通了整个业务流程，为新业务的应用提供了一个统一的基础承载面。同时，借助IPv6的基础和新型能力也为解决业务的需求和痛点问题提供了新的思路。但上述虚拟以太网技术并没有充分利用IPv6的优势，仅将IPv6和IPv4作为同类承载技术对待，因此在封装效率、安全或流量调度方面存在不足。

本文中我们提出了一种基于IPv6数据平面的承载多站点以太网虚拟专网技术方案EVN6（即基于IPv6的虚拟以太网），它将要传送的以太网数据帧直接封装在IPv6数据包中，能很好地满足云网融合技术快速发展带来的新要求，其技术特点有：

1）网络覆盖范围广泛，支持运营商间、域间的互通和协同，不存在MPLS等底层传输技术覆盖范围方面的限制。

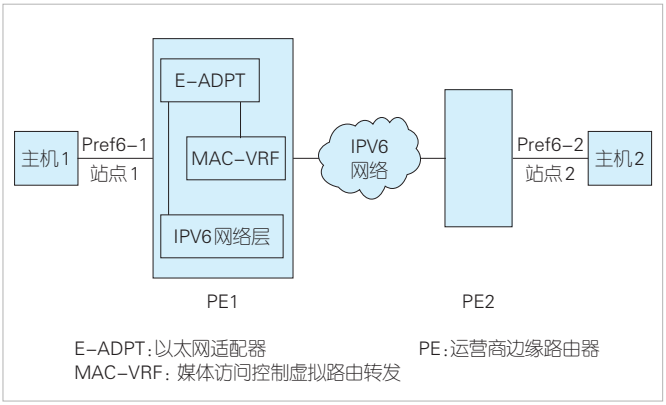
2）在安全性方面，没有采用传统隧道的静态端点地址，避免了由显式隧道端点地址而带来的分布式拒绝服务（DDoS）攻击等风险。

3）具有更高的封装效率，减少了封装的开销以及由此带来的处理开销。

4）同一站点中的不同主机具有不同的外部IPv6地址，可以根据源或目的IPv6地址实现流量负载均衡。

1 EVN6技术架构

EVN6是一种在IPv6网络中承载多站点以太网虚拟专网的方案。它将要传送的以太网数据帧直接封装在IPv6数据包中，并利用媒体访问控制（MAC）地址、虚拟网标识等信息生成特有的IPv6源和目的地址，支持将以太网帧传送到目的站点。本方案可应用于企业站点互联以及数据中心等多个场景。相关的技术方案已经转化为国际互联网工程任务组（IETF）文稿，并提交至Internet Area Working Group[6]。支持EVN6的系统架构如图1所示。



▲图1 EVN6技术架构图

EVN6在传统的路由架构下引入一个关键功能组件以太网适配器（E-ADPT），负责处理转发面的封装和路由面的路由信息学习等关键任务。其中，MAC虚拟路由转发（MAC-VRF）是一种基于MAC的虚拟路由转发表，用来存储主机MAC地址信息、IPv6前缀等相关信息以及转发规则。E-ADPT和MAC-VRF通常在运营商边缘路由器（PE）设备上实现，部署在用户站点网络的出口，承载使用虚拟以太网的业务流量。

通常，以太网虚拟网络由分布在不同地理位置的多个站点组成，每个站点都通过IPv6网络边缘的本地PE连接到IPv6网络。为了区分不同的以太网虚拟网络实例，长度为32位的虚拟以太网标识（VEI）可进行全局识别，最多可识别42.9亿个以太网虚拟网络。E-ADPT将客户站点要传输的以太网数据帧直接封装成IPv6数据包，并发送到IPv6网络。对于接收到的发往本地站点的IPv6数据包，E-ADPT会删除其数据包标头并恢复原始以太网帧。

对于一个指定的以太网虚拟网络，E-ADPT使用IPv6站点前缀，即Pref6，来标识不同的站点，因此在同一个以太网虚拟网络实例中的不同站点的Pref6是不同的。但不同的以太网虚拟网络实例中的同站点的Pref6可能是相同的。架构要求IPv6地址属于全球单播地址类型，并且可以在全局路由系统中被访问。地址可以从网络运营者的IPv6地址空间中选取，而不用再向互联网注册机构申请专用的地址块。需要注意的是，Pref6的长度可以是可灵活选择的，它可以等于或小于64位。当Pref6的长度小于64位时，Pref6通常在64位的高位，同时在低位补零。对于一个具有多个站点的以太网虚拟网络，VEI与站点前缀之间存在1∶*N*的关系。

为了将以太网帧通过IPv6网络发送到正确的目标站点，PE中的MAC-VRF表用于存储以太网虚拟网络中所有主机的MAC地址、相应以太网的VEI和站点的Pref6。MAC-VRF应至少包含MAC地址、VEI值和相应站点的Pref6相关的信息。MAC-VRF为E-ADPT转发面的封装功能提供了必要的信息，因此E-ADPT在发送数据包之前应该接收上述信息。E-ADPT控制面提供了一种机制，在站点之间自动传递MAC地址和Pref6的映射信息。当PE接收到站点内主机发出的以太网帧时，它会使用目的MAC地址作为一个索引在MAC-VRF表格中查找相对应的Pref6和VEI的值，然后用上述信息封装数据包。

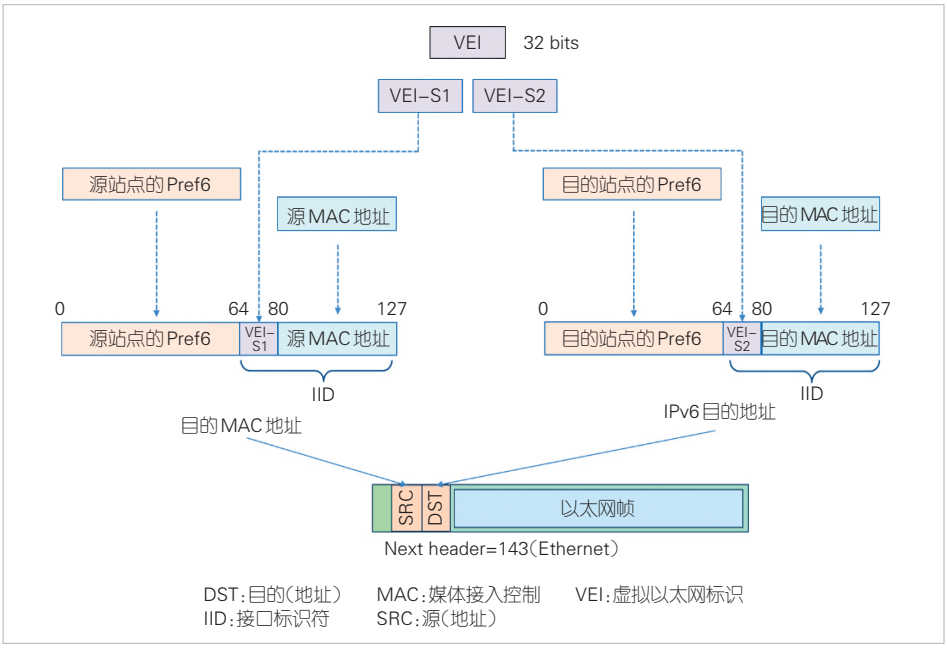
2 EVN6数据面关键技术

数据面是指实际数据包接收、处理和转发的能力集合，负责传输实际数据。EVN6针对数据包的封装和生成提出了新型技术框架，增强了数据面能力，包含编址技术、同子网

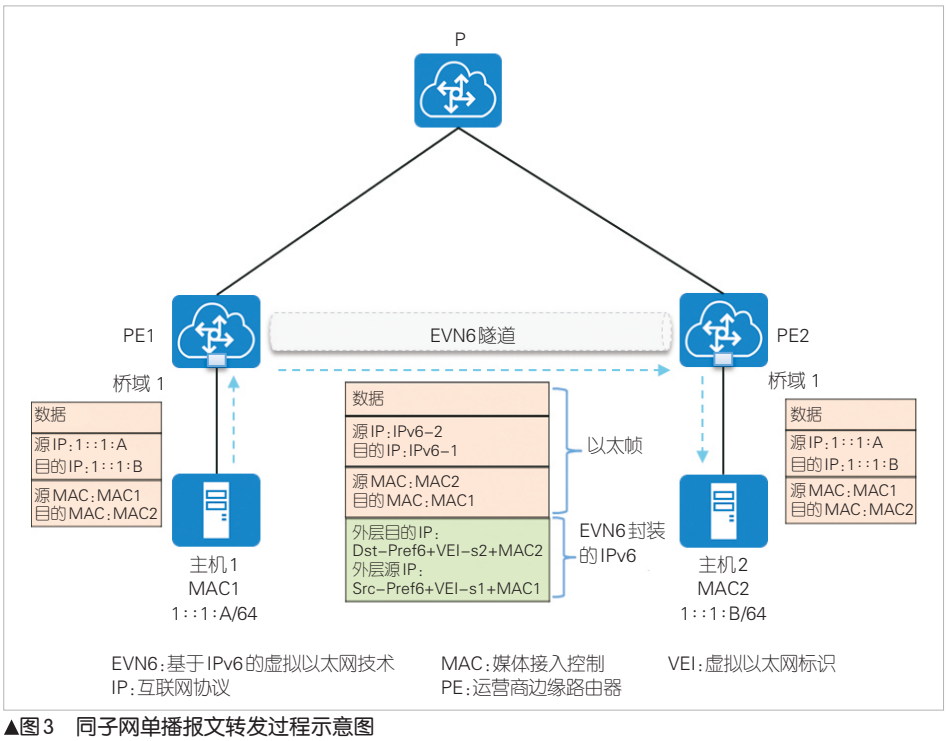
单播报文和BUM（广播、组播、未知单播报文的统称）报文转发等技术。

2.1 EVN6编址技术

EVN6在编址方面所需的关键配置信息包括VEI、IPv6前缀和通信双方主机的MAC地址。其中，VEI是32位的虚拟以太网的实例标识，IPv6前缀（Pref6）标识站点在网络中的逻辑位置也代表站点的路由可达性，MAC地址是通信双方物理或虚拟形态主机的48位以太网地址。EVN6数据包的地址架构和现网中常用的全球IPv6单播地址架构保持一致，长度为128位。其中，前64位为网络前缀，后64位为接口标识符。地址具体的编址方式如图2所示，即根据隧道两端的EVN6配置和以太网数据帧中的MAC地址，生成EVN6的IPv6报文源地址和目的地址。总体来看，前缀由Pref6构成，后缀由VEI和MAC地址映射生成。对于源地址，源站点对应的Pref6作为前64位的网络前缀，VEI的前16位（VEI-S1）和源主机的48位MAC地址连接合成后64位接口标识符。而对于目的地址，目的站点对应的Pref6作为前64位的网络前缀，VEI的后16位（VEI-S2）和目的主机的48位MAC地址连接合成后64位接口标识符。IPv6报文头部中的Next head字段值设置为143，代表以太网数据帧。值得注意的是，EVN6的地址封装通过算法完成转换，无须建立地址表记录相关连接的状态，因此是一种无状态的映射方式，减少了系统实现的复杂度和运维风险。



▲图2　EVN6数据面的编址方式



▲图3　同子网单播报文转发过程示意图

2.2 同子网单播报文转发

数据面的关键能力之一是支持同子网单播报文的转发，这也是虚拟以太网最基础的应用场景。我们以图3的网络拓扑图为例阐述其转发过程。PE1设备和PE2设备开启了EVN6功能，形成了一个简单的单点到单点的EVN6网络，其中主机1准备向主机2发送数据包。主机1生成以本身MAC1地址为源地址和以主机2的MAC2地址为目的地址的以太帧。当收到来自主机1的以太数据帧时，EVN6网络PE1将根据接入端口和桥域（BD）信息获取对应的二层虚拟以太网的实例标识VEI，在

该二层虚拟以太网内查找MAC-VRF表，获得目的MAC2地址所对应的站点前缀Dst-Pref6。PE1设备在获得上述信息后，根据2.1章节中的编址方法生成外部IP包的源和目的地址，进行IPv6数据包的封装。

PE2在收到IPv6数据包时，检查其目的地址前缀与PE2上的站点前缀是否匹配。若匹配则将地址中的VEI-s1和VEI-s2提取出来，组合成VEI并对该VEI值进行校验。之后再检查Next header的值是否为143。若是，则为EVN6以太帧封装，然后去掉EVN6封装，根据MAC表将以太帧转发给主机2；否则，该IPv6数据包则会按照丢弃的方式处理。

2.3 BUM报文转发

二层转发设备在转发报文时，转发类型只有4种：广播、组播、未知单播和已知单播报文。BUM报文是广播、组播、未知单播报文的统称，其MAC地址具备如下特点：MAC地址为全FF的报文则为广播流量；MAC地址第一个字节的最低位为1的报文则为组播流量；在MAC表中没有表项的单播报文为未知单播流量。

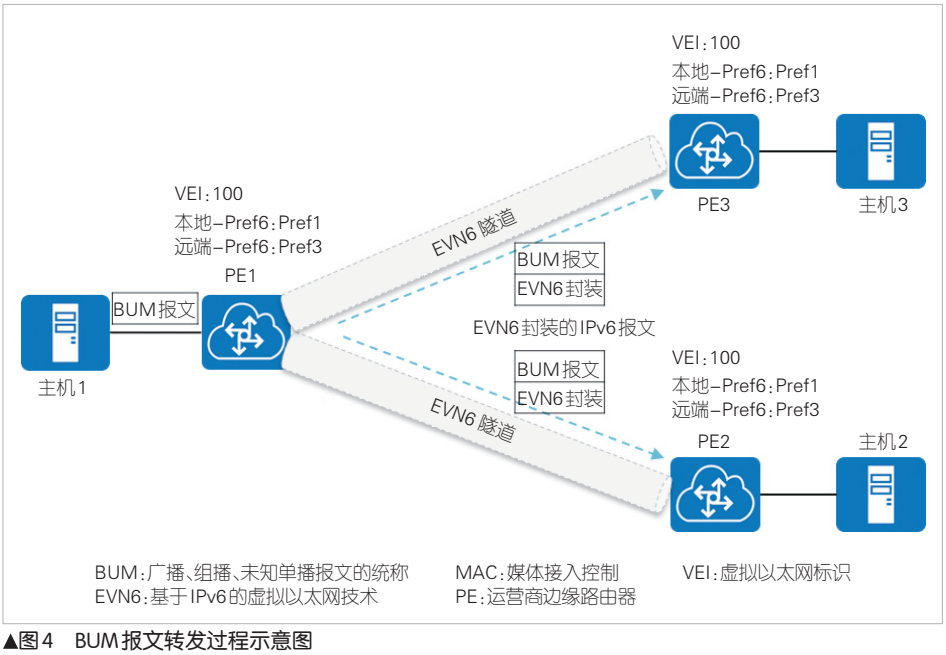
BUM报文在进行EVN6封装时，基于上述MAC地址，其目的地址映射规则如下：

广播流量：Dst-Pref6+VEI-s2+广播MAC地址（全FF）；

组播流量：Dst-Pref6+VEI-s2+组播MAC地址；

未知单播流量：Dst-Pref6+VEI-s2+单播MAC地址。

在2.2章节描述的报文转发过程是已知单播报文转发，而BUM报文的转发采用头端复制的方式，如图4所示。



▲图4　BUM报文转发过程示意图

当BUM报文进入EVN6隧道时，接入端会根据头端复制列表进行报文的EVN6封装，并将报文发送给头端复制列表中的所有出口端。这种方式可以确保BUM报文能够被正确转发到目的设备。以广播报文的转发为例，当PE1收到主机1发送的广播数据帧时，需要将该数据帧传送到所有具有相同虚拟以太网实例VEI的站点。其中，E-ADPT以VEI为索引查找MAC-VRF表中所有相关联的站点前缀，然后分别生成对应站点的报文。当PE2收到该报文时，E-ADPT提取其中的目的站点前缀和虚拟以太网实例值。若结果与本站点的Pref6和VEI值相匹配，则去掉EVN6封装并向站点内的主机转发报文。

3 EVN6控制面关键技术

控制面制定数据面策略和配置，指导数据包的操作，负责传输控制信令。控制面可实现EVN6隧道的自动创建、数据面关键信息的发现和传送等操作，降低网络部署的复杂性和扩展难度。本章将重点介绍静态隧道创建的流程，同时给出动态隧道创建方案初步设计的一些考虑。

3.1 EVN6隧道静态创建流程

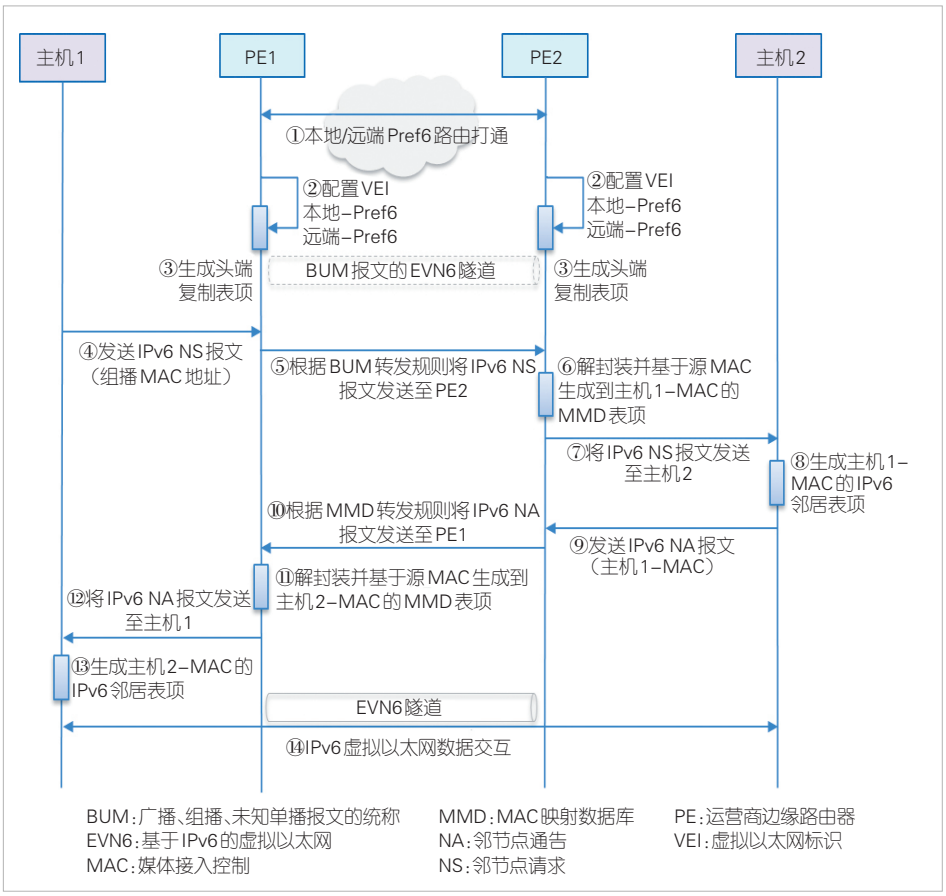
静态创建EVN6隧道就是在EVN6网络隧道两端的PE设备上手工配置站点信息、虚拟以太网信息以及两者之间的映射规则，实现端到端隧道的互通。在这种方式下，EVN6配置虚拟以太网标识VEI、本端站点的IPv6前缀和远端站点的IPv6前缀。静态配置EVN6隧道的时序图如图5所示。

关键的步骤如下：

1）通过静态路由或者动态路由协议打通站点前缀Pref6在底层网络的路由。

2）两端站点的PE配置VEI、本地Pref6和远端Pref6。如果到远端站点前缀的路由可达，则基于配置生成头端复制表项。头端复制表项主要用于BUM报文的封装和转发。MAC-VRF表中头端复制表项参考示例如表1所示。

3）Host之间通过IPv6 ND协议（MAC地址为组播的NS报文/单播的NA报文）互相学习主机MAC和IPv6地址，并形成各自的IPv6邻居表项，可在PE设备上启用邻居发现（ND）代理来减少邻居发现协议



▲图5　EVN6隧道静态创建时序图

（NDP）消息的泛洪。同时在PE的MAC-VRF表中同步生成一条主机MAC与EVN6隧道的映射关系，如表2的PE1映射表项所示。Host之间进行以太帧通信时，根据该表项进行EVN6封装。

3.2 EVN6隧道动态创建方案的考虑

EVN6隧道可以采用手工的方式静态创建，也可以采用边界网关协议（BGP）EVPN作为控制面动态创建。这种方式不仅可以实现EVN6隧道的自动建立，从而降低网络运维的复杂度和提升网络的可扩展性，而且可以实现IP、MAC、VEI和主机路由信息的自动宣告，从而有效减少了BUM洪泛流量。

EVPN是下一代全业务承载的VPN解决方案。EVPN统一了各种VPN业务的控制面，利用BGP扩展协议来传递二层或三层的可达性信息，实现了转发面和控制面的分离。EVPN逐渐演进为一套通用的控制面协议，而不是为了承载业务的数据面协议。因此，EVN6拟在EVPN中设计新的协议类型和路由类型来承载和传递控制面中的信息，包括L2 VPN地址族标识（AFI），EVPN子地址族标识（SAFI）和EVPN网络层可达性信息（NLRI）。控制面的具体方案还在设计之中，不在本文中赘述[7]。

4 EVN6封装效率分析及现网技术验证

在EVN6方案中，数据包的封装方式得到了进一步简化。传送的以太网数据帧被直接放置在IPv6数据包的净荷中，在保障网络标识和路由能力的前提下，取消了VXLAN技术存在的多层封装，显著降低了封装开销和多协议层处理数据的成本。如图6所示，在VXLAN封装方式下，以太网数据帧先被分别封装在8字节的VXLAN报头、用户数据报协议（UDP）报头中，最后封装在IPv6报文之中，总共需要

▼表1　头端复制表项

VEI	IPv6映射前缀（本地Pref6）	IPv6映射前缀（远端Pref6）	主机MAC	主机IP	FLAG
100	Pref1	Pref2	00:00:00:00:00:00	——	BUM
100	Pref1	Pref3	00:00:00:00:00:00	——	BUM

BUM：广播、组播、未知单播报文的统称　FLAG：标志　MAC：媒体接入控制　VEI：虚拟以太网标识

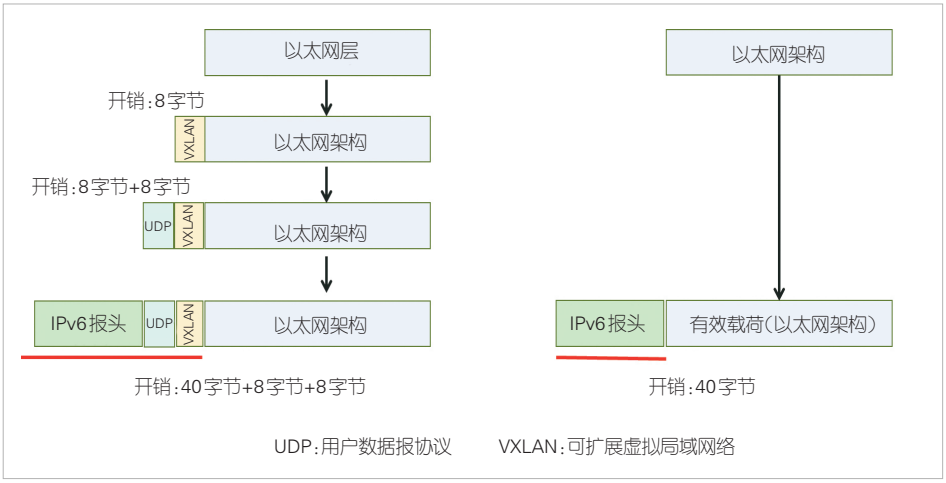
▼表2　MAC与EVN6隧道映射表项

VEI	IPv6映射前缀（本地Pref6）	IPv6映射前缀（远端Pref6）	主机MAC	主机IP	FLAG
100	Pref1	Pref2	00:00:00:00:00:00	——	BUM
100	Pref1	Pref2	主机2-MAC	主机2-IP	

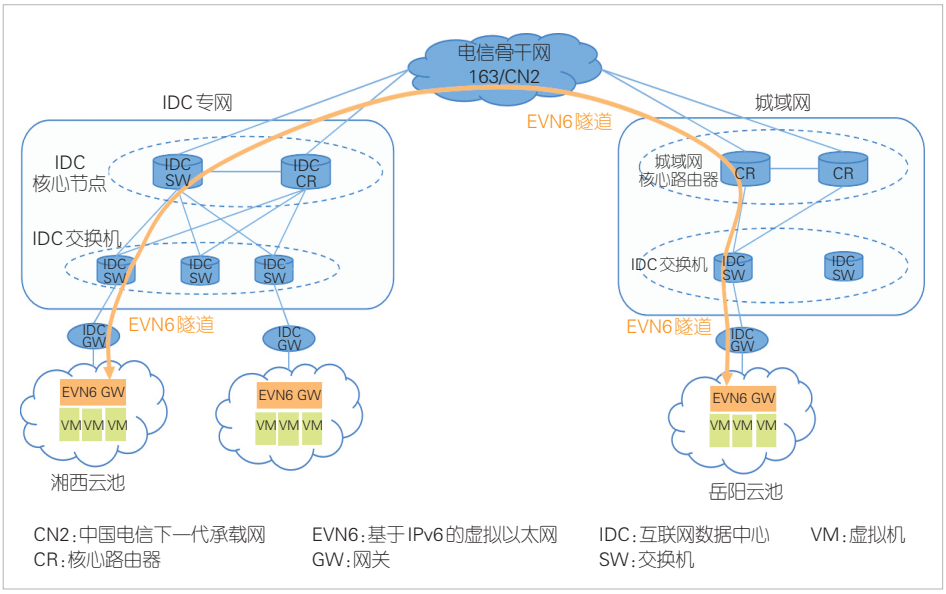
BUM：广播、组播、未知单播报文的统称　FLAG：标志　IP：互联网协议　MAC：媒体接入控制　VEI：虚拟以太网标识

56字节。而在EVN6封装方式下，以太网数据帧直接封装在IPv6报文中，报文长度减少了16字节，整体降低了28.5%的封装开销。

针对云间跨域互联场景，现网试验采用了静态EVN6方案，网络拓扑图如图7所示。该方案测试EVN6隧道的创建、EVN6转发面封装的正确性，以及基于EVN6虚拟以太专网的业务互通等功能。该虚拟以太网在两个地市云资源池之间建立传输网络，需要跨域骨干网和城域网，并且要求在底层网络和云资源池全面支持IPv6。为了建立EVN6网络，需要分别在两个资源池上虚拟一台主机作为网关部署EVN6镜像软件，另外创建两个虚拟机部署云上的应用，在试验中发起业务数据流。虚拟机要求必须支持IPv6，两台网关配置站点IPv6前缀，与网关相连接的虚拟机配置同一IPv6子网地址。云间互联网络需要打通两个站点前缀的路由，确保业务数据能够通过IPv6的方式转发到对端网关。



▲图6 EVN6和VXLAN封装方式开销对比



▲图7 EVN6技术试验组网

EVN6试验系统上线后，租户网络之间实现了业务隔离，云间的虚拟机迁移等业务运行良好，满足了云间基于IPv6的虚拟以太网互联业务的基本需求。本次试验展现了如下3个方面的效果：

1）EVN6业务可以实现快速开通。当前运营商网络基本全面支持并开启了IPv6协议栈，目前省内城域网到163骨干网有Tbit/s级别的预留带宽，因此整个业务发放只需要配置首尾网关两端，中间节点不需要任何重新的规划和配置改动，也不涉及跨域的设置，从而达到业务快速开通的目的。

2）EVN6与IPv6网络具有良好的兼容性，可以非常方便地进入数据中心、城域网和骨干网等网络，并且可以跨越多个自治域组网，几乎不存在覆盖范围的限制。其他协议如SRv6、MPLS等则存在“有限域”组网方面的限制。此外，EVN6是在IPv6基本报头做的技术创新，和SRv6等IPv6+技术也是兼容的，可以结合IPv6的新型技术，从而充分发挥路径编程等IPv6能力。

3）EVN6利用IPv6海量地址空间的特性，在三层编址空间融入了二层以太网、主机等信息，提供了一个在网络层根据业务网络信息转发和调度流量的接口，可通过更简洁的方式提供流量工程和负载均衡，更好地满足智慧化运营要求。

5 结束语

EVN6是一种基于IPv6的新型虚拟以太网组网技术，该技术简化了数据面的封装方式，基于IPv6协议直接承载虚拟以太网数据包，提高了封装效率。利用了IPv6海量的地址空间，在编址层面融入丰富的信息，充分发挥了IPv6的基础能力。EVN6技术标准的进一步成熟和应用将会对

L2VPN业务领域产生创新驱动作用，带动相关产业链的升级，同时也符合国家大力推动IPv6规模部署应用工作的要求，进一步释放IPv6技术潜力。未来EVN6有潜力作为一种简化的L2VPN方案广泛应用于运营商网络。

致谢

感谢清华大学李星教授、包丛笑教授对本研究的帮助！

参考文献

[1] IETF. BGP/MPLS IP virtual private networks (VPNs): RFC 4364 [S]. 2006
[2] IETF. Virtual extensible local area network (VXLAN): a framework for overlaying virtualized layer 2 networks over layer 3 networks: RFC 7348 [S]. 2014
[3] IETF. BGP overlay services based on segment routing over IPv6 (SRv6): RFC 9252 [S]. 2022
[4] 解冲锋, 李星, 李震, 等. 大规模网络向IPv6单栈演进的技术方案 [J]. 中兴通讯技术, 2022, 28(1): 57-61. DOI: 10.12142/ZTETJ.202201012
[5] 马晨昊, 解冲锋, 郑伟, 等. 5G SA网络引入IPv6的思路探讨 [J]. 中兴通讯技术, 2020, 26(3): 43-48. DOI: 10.12142/ZTETJ.202003009
[6] XIE C F, LI X, BAO C X, et al. EVN6: a framaework of mapping of ethernet virtual network to IPv6 underlay, draft-xls-intarea-evn6-00 [Z]. 2024
[7] XIE C F, SUN J B, LI X, et al. EVPN route types and procedures for EVN6, draft-xie-bess-evpn-extension-evn6-00 [Z]. 2024

作 者 简 介

马晨昊，中国电信研究院网络技术研究所工程师，现任ETSI TC INT副主席；主要从事IPv6、“IPv6+”、未来网络等相关技术的研究，以及协议一致性和互通性测试标准化工作；曾获得中国通信标准化协会科学技术奖一等奖等。

孙吉斌，中国电信研究院网络技术研究所工程师；主要研究领域为未来网络关键技术、IPv6、算力网络等；参与了EVN6相关标准研究和原型系统研发工作。

解冲锋，中国电信研究院集团级高级技术专家，教授级高工，中国通信学会会士，中国互联网协会学术委员会副主任委员，北京市IPv6重点实验室主任，曾在美国UCLA大学做政府公派访问学者一年；长期从事宽带网络架构、IPv6下一代互联网、物联网、网络安全、云网融合等方面的研究；参与制定国家IETF RFC标准6项，曾获得2023年度国家科技进步奖一等奖和2023年度中国通信标准化协会科学技术奖一等奖，2019年获得“政府特殊津贴”。

广域抗损高吞吐URDMA技术

URDMA Technologies for Wide-Area High-Throughput Network

段晓东/DUAN Xiaodong，陆璐/LU Lu，孙滔/SUN Tao，
李志强/LI Zhiqiang，杨红伟/YANG Hongwei，
杜宗鹏/DU Zongpeng

（中国移动通信有限公司研究院，中国 北京 100053）
(China Mobile Research Institute, Beijing 100053, China)

DOI：10.12142/ZTETJ.202406005
网络出版地址：http://kns.cnki.net/kcms/detail/34.1228.TN.20250121.1330.002.html
网络出版日期：2025-01-21
收稿日期：2024-10-12

摘要：随着国家“东数西算”战略实施以及智算、超算业务的快速发展，海量数据广域传输需求不断增多。提出一种广域抗损高吞吐超远程直接内存访问（URDMA）技术方案，通过对传输控制协议/互联网协议（TCP/IP）协议栈的完全卸载，消除中央处理器（CPU）对网络高吞吐性能的限制。采用拥塞控制、丢包恢复、丢包重传等技术增强标准第2代基于融合以太网的远程直接内存访问（RoCEv2）协议，使其在广域有损网络下保持高吞吐性能。测试结果表明，在往返时延（RTT）时延为20 ms、丢包率0.1%的网络环境下，TCP协议吞吐性能仅为0.02 Gbit/s，标准RoCEv2性能接近为0，URDMA协议吞吐性能为88.26 Gbit/s；当RTT时延增加到80 ms时，TCP和RoCEv2协议吞吐基本衰减为0，URDMA协议吞吐性能为83.12 Gbit/s，仍然保持较高的性能。

关键词：广域抗损高吞吐；数据快递；远程直接内存访问；RoCEv2

Abstract: With the implementation of the national "East Data West Computing" strategy and the rapid development of intelligent computing and supercomputing services, the demand for large-scale data transmission is constantly increasing. A wide-area high-throughput ultra remote direct memory access (URDMA) technology solution is proposed, which mitigates the limitation of the central processing unit (CPU) on high-throughput network performance by completely offloading the transmission control protocol/Internet protocol (TCP/IP) protocol stack. By adopting congestion control, packet loss recovery, packet loss retransmission, and other technologies to enhance the 2nd version of remote direct memory access based on converged ethernet (RoCEv2) protocol, URDMA enables high-throughput performance in wide-area lossy networks. The test results show that in a network environment with a round-trip time (RTT) of 20 ms and a packet loss rate of 0.1%, the TCP protocol throughput performance is only 0.02 Gbit/s, the standard RoCEv2 performance is close to 0, and the URDMA protocol throughput performance is 88.26 Gbit/s. When the RTT increases to 80 ms, the TCP and RoCEv2 protocols basically decay to 0, and the URDMA protocol throughput performance is 83.12 Gbit/s, still maintaining high performance.

Keywords: high-throughput in wide-area network; data express; remote direct memory access; RoCEv2

引用格式：段晓东, 陆璐, 孙滔, 等. 广域抗损高吞吐URDMA技术 [J]. 中兴通讯技术, 2024, 30(6): 23-30. DOI: 10.12142/ZTETJ.202406005
Citation: DUAN X D, LU L, SUN T, et al. URDMA technologies for wide-area high-throughput network [J]. ZTE technology journal, 2024, 30(6): 23-30. DOI: 10.12142/ZTETJ.202406005

2022年国家发展和改革委员会、中央网信办、工业和信息化部、国家能源局联合启动了“东数西算”战略。随着东数西算战略的实施，东数西存、东数西训、东数西渲等场景对海量数据跨广域网数据快递需求日益凸显。随着产业数字化、云计算、分布式人工智能（AI）的发展，数据异地上云、云迁移、云灾备、跨智算中心互联等时空大尺度数据搬迁场景中数据规模越来越大，对网络吞吐的要求越来越高。

网络带宽也从10G发展到25G、100G、200G、400G、800G甚至1.6T。与网络带宽快速增长形成鲜明对比的是，后摩尔时代中央处理器（CPU）算力增速远低于网络带宽增速，并且差距还在持续增大。如何充分利用网络带宽破解海量数据广域传输瓶颈，如何以低算力损耗满足高速网络处理和传输要求，对立体泛在算力网络整体算效提升及分布式AI训练、推理性能提升至关重要。

本文中，我们将重点分析数据快递业务对广域网络的高吞吐需求与挑战，并给出数据快递广域抗损高吞吐超远程直接内存访问（URDMA）解决方案及其初步测试结果。

1 数据快递广域高效传输需求与挑战

1.1 数据快递广域高效传输需求

东数西算、数据异地上云、云间灾备、广域智算互联等场景大多涉及海量数据跨省传输。自动驾驶数据上云需要传

输大量数据至智算中心进行训练，每辆车每天生成的数据量可达几TB至十几TB，完成L3级别的训练可能会产生8 EB的数据；天文数据计算，FAST每年约200多个[1]观测项目，单项目产生观测数据量TB至PB量级，年产数据约15 PB。为了缩短数据传输的时间，需要借助高带宽网络。但高带宽不等于高（有效）吞吐，端到端高吞吐才是确保数据时效性和减少传输成本的关键。在面向连接的可靠传输技术中，距离越长，确认报文回复耗时越长，对业务发送端和接收端服务器的缓存要求越高，实现精确丢包检测的难度越大，实现长距离、高吞吐的可靠传输挑战越大。

摩尔定律放缓使得通用CPU性能增长的边际成本迅速上升。数据表明，现在CPU的性能年化增长（面积归一化之后）仅有3%左右[2]，这导致带宽性能增速比（RBP）失调。网络的带宽年化增长在2010年前大约是30%，2015年微增到35%，近年达到45%。相应地，CPU的性能增长从10年前的23%下降到12%，并在近年直接降低到3%。在这3个时间段内，RBP指标从1左右上升到3，并在近年超过了10。网络带宽的剧增对业务发送端和接收端服务器的数据收发处理能力提出了更高要求，基于CPU的传输控制协议/互联网协议（TCP/IP）等传统处理方式逐渐成为端到端高吞吐数据传输的瓶颈。

1.2 数据快递广域高效传输挑战

长距高吞吐是数据快递业务的重要目标。2020年中国移动完成了全球最大的云专网商用部署，基于软件定义网络（SDN）和段路由（SR）技术的云专网实现跨数据中心云资源池的整合，骨干段覆盖全国所有直辖市、省、自治区；在云骨干网中，网络物理带宽已不是瓶颈，如何提升端到端高有效吞成为关键。目前主流业务多采用TCP/IP协议进行海量数据广域传输。由于现有传输协议、拥塞控制算法、丢包冗余恢复机制、选择性重传机制及算力损耗等方面的限制，现有协议和机制无法满足“长肥”网络下的高吞吐数据传输需求。

1）协议。互联网工程任务组（IETF）RFC1323[3]标准规定，TCP理论窗口最大值为1 GB（2^{30} bytes）。依据包守恒定律（理想情况下的窗口大小和Inflight数据相同），当吞吐量为400 Gbit/s时，单流最远传输约为1 000 km；当吞吐量为800 Gbit/s时，最远传输仅500 km，无法满足广域跨智算中心分布式AI训练等少数大象流高吞吐长距离传输需求。

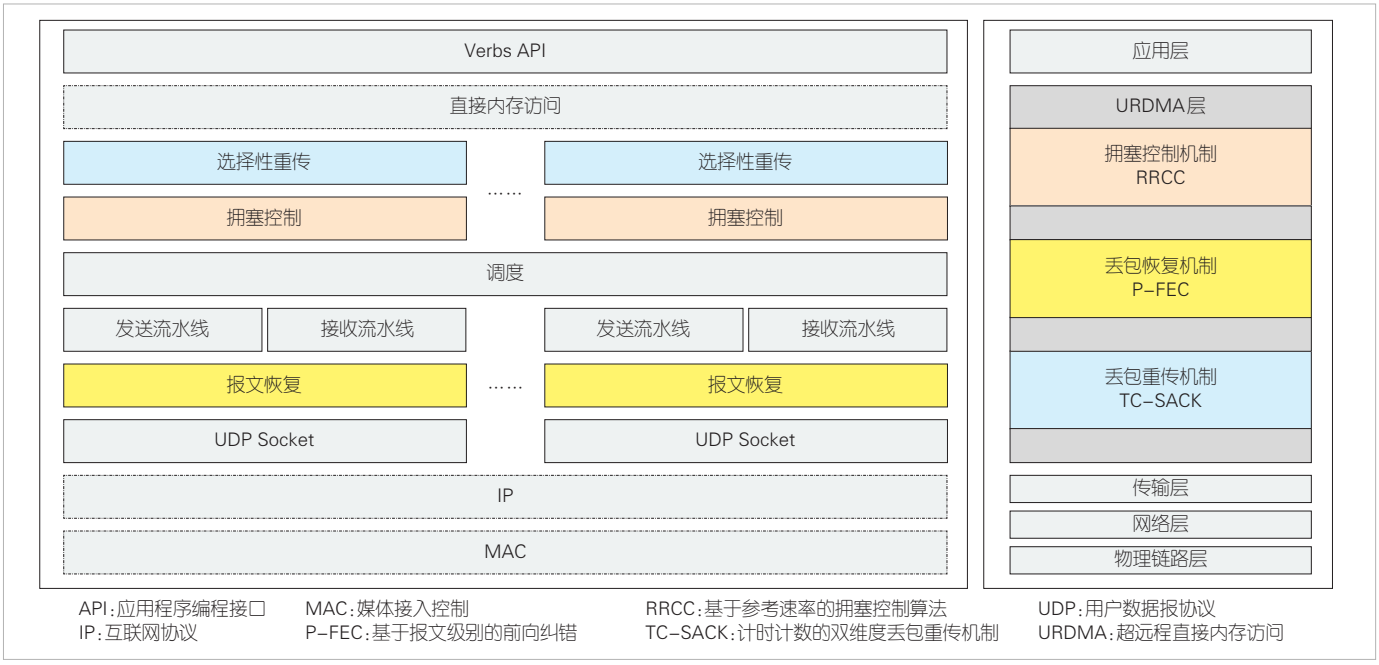
2）拥塞控制算法。拥塞控制算法按照拥塞判断依据大致分为丢包类、时延类和带宽类。丢包类算法如Reno[4]、CUBIC[5]等，依据网络是否丢包来判断拥塞，但因易误判，发送速率会出现过度调整，从而限制了吞吐；时延类算法如FAST[6]、Vegas[7]等，依据网络环回时延的变化来判断拥塞，但可能因为时延突变或回路时延变大造成拥塞误判，导致发送速率过度调整；带宽类算法如瓶颈带宽和往返传输时延（BBR）[8]等，通过检测可用带宽来调整发送速率，但公平性较差，并可能因为发送速率调整不及时而导致大量丢包。

3）丢包冗余恢复机制。丢包冗余恢复机制常采用纠错码。纠错码是一种编码技术，通过在数据传输过程中添加冗余信息来保护数据的完整性。通过选择合适的算法和参数（例如，块大小、纠错码长度、冗余度等），该机制能够抵御多个数据包的丢失或损坏。现有基于纠错码的丢包恢复方法存在不足，例如，接收端在发现丢包后才通知发送端进行纠删码编码和校验块发送，接收端必须等待其所请求的校验块被完全接收后，才能实现丢失恢复。这会产生极大的额外传输时延。对丢包的判断、对数据的编码和对数据的解码均需CPU持续参与控制，会占用大量的CPU时间，从而产生额外的CPU负担。

4）选择性重传机制。TCP有超时重传（RTO）和快速重传两种机制。RTO根据往返时延（RTT）来估算，广域高RTT环境等待时间长、效率低；快速重传在收到的重复的确认字符（ACK）达到3个之后就进行重传，效率高，但需要充足的后续报文。TCP/IP协议目前都依赖发送端的流量控制，发送端通过ACK推测网络的情况，发送速率严重依赖接收端的ACK反馈，不利于广域长距离网络的高吞吐传输。

5）算力低损耗。传统的TCP/IP协议栈运行在操作系统内核空间，主机间通信需要用户空间、内核空间、硬件网卡之间多次交互[9]。数据拷贝、用户态和内核态的切换、数据包收发中断响应等都需要CPU参与，大量的CPU资源被消耗。处理10G网络数据包需要大约4个Xeon CPU核，即仅是网络数据包收发处理就占用了一个通用8核CPU一半的算力[10]。当网络带宽增大到100G、200G甚至400G时，CPU性能将会成为高吞吐传输的瓶颈。因此出现了很多卸载方案，这些方案将协议栈的全部或部分数据包处理工作卸载到硬件网卡上。这样可以充分发挥硬件网卡对数据包的高速处理能力，降低CPU损耗，提升数据收发端（例如业务服务器）的数据包处理能力。其中有代表性的包括TCP/IP协议卸载引擎（TOE）[11]、数据面开发套件（DPDK）[12]，但卸载效果都不是很理想。

RDMA使用内存零拷贝、内核旁路、CPU卸载等技术，将协议栈全卸载到网卡处理，允许用户态的应用程序直接读取和写入远程主机内存，避免了数据拷贝和上下文切换，实现了高吞吐量、低时延和低CPU算力损耗。RDMA有3种技术路径：InfiniBand[13]、基于融合以太网的RDMA



▲图1 URDMA技术架构

（RoCE）[14]和互联网广域RDMA协议（iWARP）[15]。其中，RoCEv2[16]因其兼容传统TCP/IP协议、易于部署管理等优点，在数据中心网络广泛应用。RDMA协议对丢包容忍度较低，要求在无损网络环境中运行，1%的丢包会使其吞吐下降至0。广域网难以实现真正的无损。

2 URDMA架构与关键技术

针对TCP/IP及标准RDMA在广域高效传输场景面临的问题，我们提出了数据快递广域抗损高吞吐URDMA解决方案：包含广域抗损高吞吐协议如URDMA、反向快启动速率控制机制如基于参考速率的拥塞控制算法（RRCC）、数据块丢包恢复机制如基于报文级别的前向纠错（P-FEC）及收发解耦多维重传机制如计时计数的双维度丢包重传机制（TC-SACK）。下文中，我们首先给出URDMA的整体架构，并针对URDMA协议及拥塞控制、丢包恢复、丢包重传3个方面的关键创新展开介绍。

2.1 URDMA技术架构

URDMA架构的设计充分考虑了兼容与平滑演进。一方面，该架构中的Verbs应用程序编程接口（API）与标准RDMA保持一致，便于存量应用平滑迁移；另一方面，该架构与现有TCP/IP协议簇兼容，仅对标准RDMA传输层协议进行增强，避免对广域网中网络设备进行升级改造，降低方案部署门槛。该方案涉及以下关键技术：

1）1套URDMA协议。扩展RoCEv2报文，支持RTT内生测量，为RRCC提供精准网络状态；支持精准内存地址投递机制，逐包携带含内存地址信息的扩展传输头（RETH），为TC-SACK直接“落存”提供访问内存的虚拟地址信息。

2）3个创新机制。反向快启动速率控制机制如RRCC，通过快速拥塞发现以及发送速率调节机制，确保高吞吐传输；数据块切分的丢包恢复机制如P-FEC利用前向纠错机制，实现数据包冗余度、带宽利用率与丢包恢复精度之间的综合最优；收发解耦多维重传机制如TC-SACK，通过优化发送和接收端滑动窗口大小和重传阈值，精确判断丢包，实现数据包选择重传。

2.2 URDMA关键技术

本章节中，我们对URDMA协议及创新机制展开介绍。

2.2.1 URDMA

URDMA协议的设计目标是基于标准RoCEv2协议，在高带宽时延积（BDP）广域网环境下，实现高吞吐性能的同时减少CPU算力消耗。其核心设计原则如下：

1）全卸载协议处理，吞吐性能和CPU利用率无关；

2）极简协议设计，高载荷比报文格式，状态少，易于硬件实现。

URDMA对标准RoCEv2的基础传输头（BTH）、确认扩展传输头（AETH）进行扩展，对RETH及BTH头的A字段的使用方式进行重新约束。

RTT的精度决定RRCC的反应灵敏度，因此URDMA协

议增强了对RTT精确测量的支持力度。针对BTH头的扩展，两个预留字段分别用来指示时间戳信息或时间戳信息在payload中的偏移起始位置。如果预留字段用来指示偏移起始位置，T_1和T_r的最高bit置为1，否则置为0。在payload中的时间戳长度为32 bit。其中T_1为报文离开发送端网卡时间，T_r为接收端到发送端的单向时延。针对AETH头的扩展，URDMA协议增加T_3、T_S、包序号（PSN）3个扩展字段，其中T_3为报文离开接收端网卡时间，T_S为发送端到接收端的单向时延，BTH头的A字段用来触发接收端反馈携带AETH头的ACK报文，发送端可根据RRCC等机制按需对A字段进行置位，触发对RTT的精确测量。

TC-SACK等机制需要报文在接收端直接“落存”。按照标准RDMA Write Request等操作机制，每条队列对（QP）连接的首包携带RETH内存地址信息，报文一旦发生首包丢失或乱序，后续报文将无法正确找到接收端存放此报文的虚拟内存地址，进而无法写入正确的内存物理位置。URDMA协议中采用逐包携带RETH头的方式，确保每个报文都能直接写入接收端内存正确物理位置。

2.2.2 RRCC

URDMA利用多参数协同判断拥塞状态，不断探测链路瓶颈带宽和时延，通过瓶颈链路带宽时延积精确计算和调整发送窗口大小，使在飞数据包维持在合理的范围内，从而在确保最大传输速率的同时减少网络传输的排队延迟，保证数据广域传输具有高吞吐和低丢包的性能。瓶颈带宽指的是端到端传输的网络路径上速率最慢的那段链路的带宽，该带宽决定了端到端传输的带宽上限。测量时延的目的是得到网络路径的最小RTT，即光/电信号从发端到收端的最小时延，具体大小取决于物理距离。

如图2所示，用于广域拥塞控制的RRCC具体实现如下：

1）QP之间建立可靠连接（RC），测量该链路的RTT作为初始最小RTT，然后进入启动阶段，此时发送速率呈指数增加。

2）当发送速率达到或超过瓶颈带宽时，降低发送速率，并排空缓存队列。

3）进入瓶颈带宽探测周期（每5～10个RTT为1个周期），发送速率稳定在瓶颈带宽，并周期性探测当前链路的瓶颈带宽。如果该链路的瓶颈带宽出现变化，则下一个周期的发送速率随之变化。

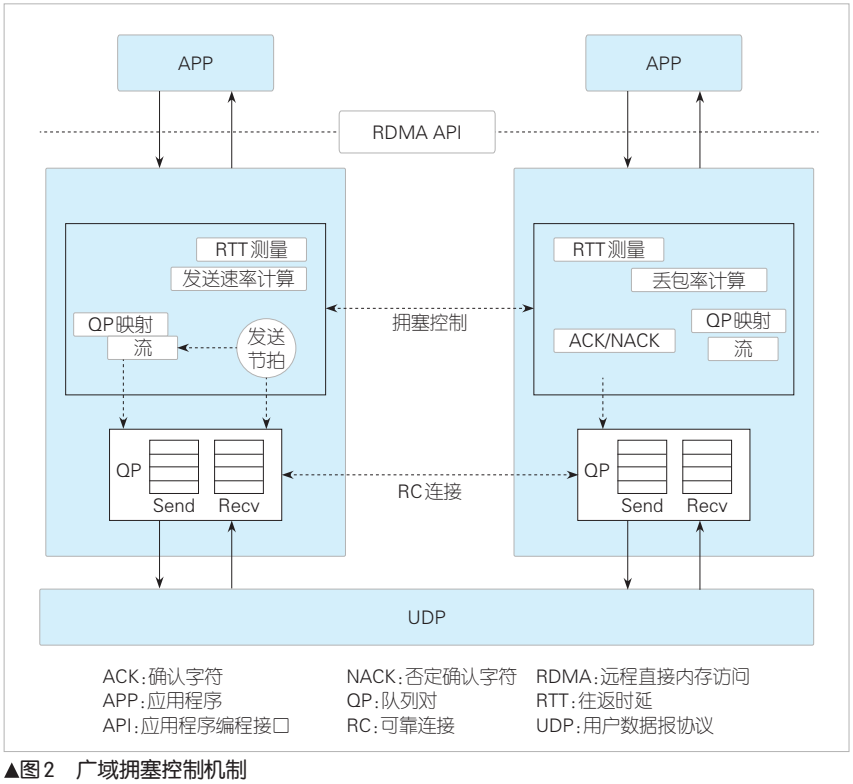
4）当接收端检测到丢包时，接收端启动丢包率统计，并将结果反馈给发送端，发送端判断丢包率大于丢包率阈值，则按比例降低发送速率。

5）设定一个固定周期，将发送速率降低至适当大小，用于测量当前链路的最小RTT。

为了进一步提升传输效率，在受控网络如中国移动云专网中，URMDA的拥塞控制机制还可以支持按照规划的参考速率进行传输，以降低端侧对广域网复杂网络情况的误判率。

在RRCC机制中，广域网的网络带宽资源将在数据中心（DC）出口路由器进行规划和分配，在数据快递的流量启动传输之前，可以从管控模块得到一个参考速率进行传输。这个参考速率可以是基于历史统计的可用带宽，或基于专线规划的带宽。一方面，端侧可以省略慢启动的带宽探测操作，直接按照参考速率开始传输；另一方面，在拥塞控制算法判定需要降速时，可以将测量到的新的瓶颈带宽与参考速率比较后取最大值作为瓶颈带宽，即锁定实际发送速率的下限。

在数据快递的场景中，相关的流通常是大象流。大象流持续时间长，带宽需求较大，但是流数较少，且部分流量可以容忍一定的发送延时如隔日达等。受控网络可根据晚上的网络负载历史数据，在网络负载较低时启动RRCC传输机制，以较快速率发送数据快递流量，达到削峰填谷的效果。



▲图2 广域拥塞控制机制

2.2.3 P–FEC

在丢包恢复机制中，接收端在发生少量丢包时，通过发送端发送的冗余数据实现快速包恢复，从而可以减少丢包重传，降低重传时延。URDMA使用基于前向纠错码（FEC）[17-18]的丢包恢复技术。

如图3所示，发送端根据RDMA协议栈发出的原始数据包进行前向纠错编码，产生冗余修复数据包并随原始数据包一同发送至接收端。

接收端根据RDMA原始数据包头部的PSN字段判断原始数据包是否发送丢失，并通过冗余修复数据包即时地恢复，最终传输至RDMA协议栈模块。

在RDMA协议栈的数据传输过程中，同一数据流内的所有RDMA原始数据包的PSN，在没有发生丢失的情况下，保持连续递增。为了有效地进行冗余修复数据包的生成，发送端采用按分组的冗余计算方式，将数据包根据其PSN分组。分组数据包的数量可根据网络丢包率动态调整，在低丢包率网络中，减少每组数据包数量，降低发送端和接收端编解码资源占用和恢复时长；在高丢包率网络中，增加每组的数据数量，提升丢包恢复成功率。这一机制有助于接收端对原始数据包和冗余修复数据包进行有序组织和识别，以维护数据传输的完整性和可靠性。

每个分组包含连续的m个RDMA原始数据包，而每k个原始数据包生成一个相应的冗余数据包。在传输过程中，这m个原始数据包按照它们的PSN序号经过RDMA协议栈和网卡递增地发送。在前向纠错编码阶段，发送端采用数据包级别的异或运算，对分组内的每个数据包进行操作，以创建冗余修复数据包。

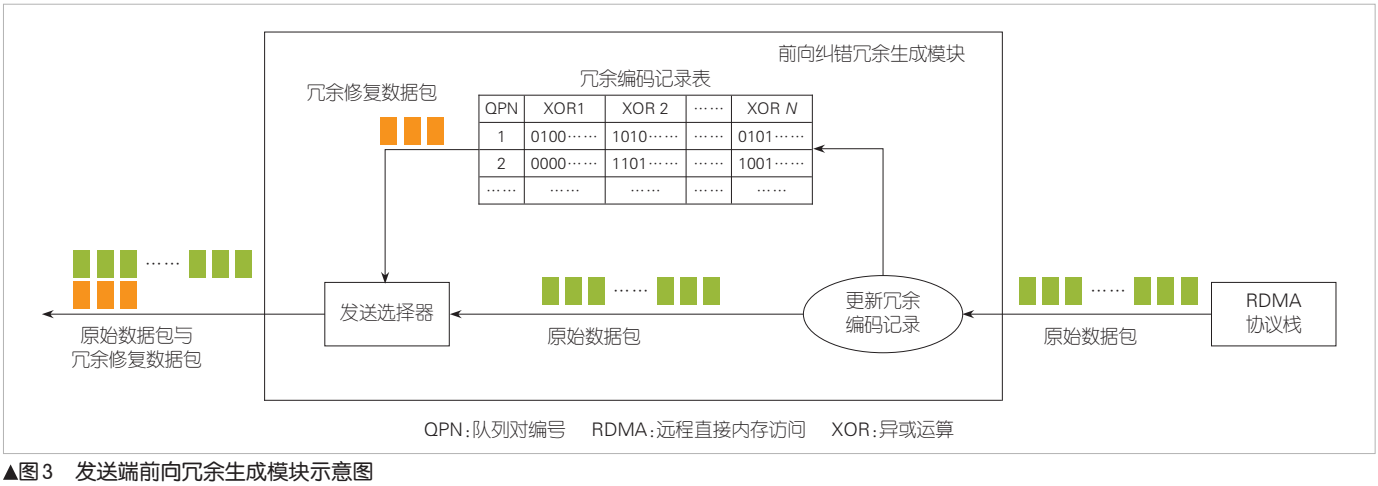
由于异或运算的属性，异或运算可以逐步执行，而无须在运算过程中记录参与异或运算的RDMA原始数据包。具体来说，每个RDMA原始数据包都可以根据其PSN确定其在分组内的位置，进而确定与之对应的编码冗余数据包，即第i个RDMA原始数据包DATA（i）对应的冗余修复数据包必然为XOR（$i \bmod r$）。因此，要计算每个冗余修复数据包的最终结果，只需要在发送第i个RDMA原始数据包时，让其与XOR（$i \bmod r$）执行异或运算即可。当该分组最后一个RDMA原始数据包发送完成时，所得到的异或计算值即所要求的冗余修复数据包。

接收端的解码原理与发送端的编码过程保持一致，都充分利用异或计算的可分步性质，将异或冗余的修复计算分散到每次单独的数据包接收过程中，如图4所示。

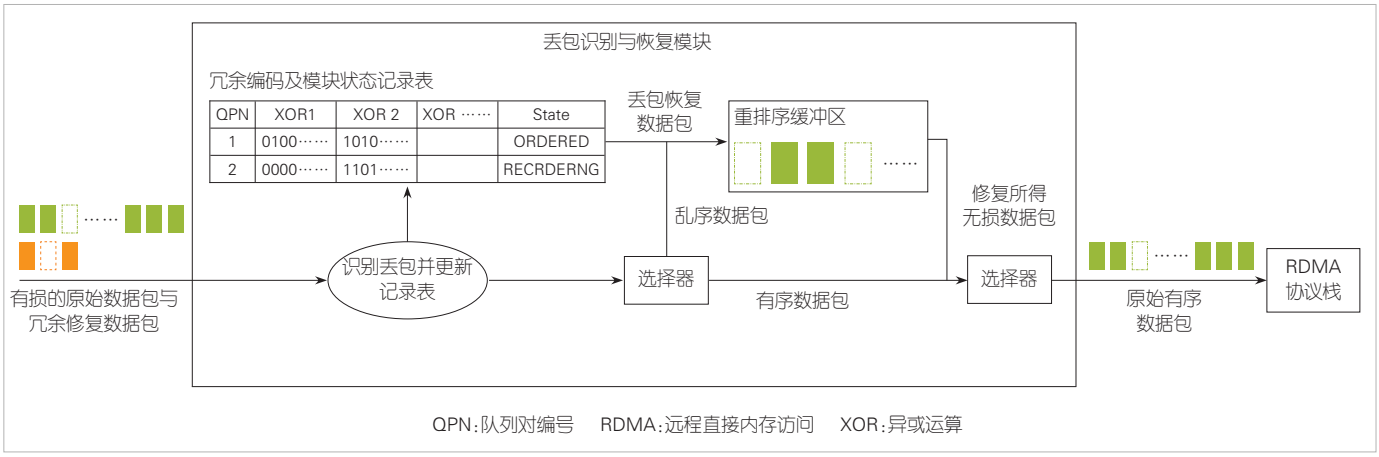
2.2.4 TC–SACK

广域拥塞控制机制能降低丢包，提升网络吞吐，为RDMA协议的运行提供良好的通路。然而，广域网无法在任意情况下都实现0丢包。丢包恢复技术仅能恢复突发性少量丢包，如果丢包数超出恢复范围，就需要启动重传机制。RDMA协议默认采用Go–Back–N丢包重传机制，如图5所示。当接收端检测到丢包时，接收端通知发送端从该丢失包之后的所有包都需要重传，即使有些包已经送达接收端。这种机制好处是接收端不需要缓存数据包，节省接收端的存储空间，且减少乱序重排的时间。但在广域有损网络条件下，Go–Back–N机制对吞吐限制巨大。

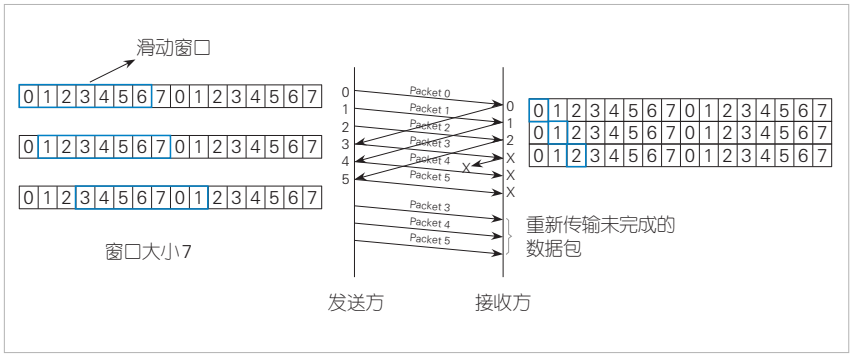
针对数据快递的高吞吐需求，URDMA提出TC–SACK机制，其主要特点为：a)无须维护接收窗口，接收端收到报文后将其直接内存访问（DMA）到内存；b)发送速率与丢包重传解耦，接收端依据预设时间或数据包数量，触发丢包重传通知。丢包信息携带在TC–SACK的ACK报文中，需要扩展扩展传输头（ETH）以支持携带丢包信息。TC–SACK的启用需要



▲图3　发送端前向冗余生成模块示意图



▲图4 接收端丢包识别与恢复模块示意图



▲图5 Go-Back-N丢包重传

在连接建立时进行协商。

在TC-SACK中，达到预设时间或数据包数量时，接收方触发确认机制。在图6中，7个数据包回复1个ACK，该ACK携带未收到的数据包的PSN，其他的乱序或顺序收到的数据包直接DMA到内存。具体流程如下：

1）发送端按照规划的速率发送数据包，接收端收到PSN=1的数据包，直接DMA到内存。这时接收端启动计时器开始计时，同时记录数据包数量为1、丢包数量为0。

2）接收端收到PSN=3的数据包，直接DMA到内存，相关的数据包需要携带目的内存地址。接收端发现PSN=2的数据包丢失，并不立刻反馈ACK。接收端记录数据包数量为3，丢包数量为1。

3）接收端收到PSN=6和PSN=7的数据包，操作类似，接收端记录数据包数量为7，丢包数量为3。到达预设的数据包数量，触发ACK，其中携带PSN2/4/5。同时接收端重置计数器和数据包计数。

4）发送端重发PSN=2/4/5数据包。

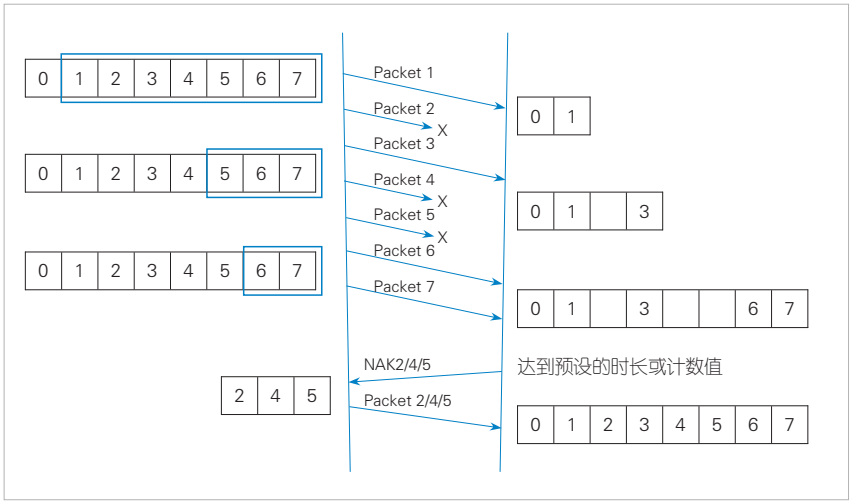
3 性能评估

3.1 测试配置

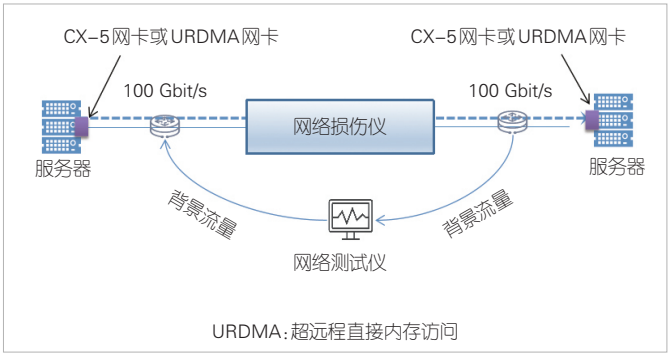
为验证URDMA网卡系统性能，我们搭建了如图7所示的测试环境，对标准RoCEv2技术、TCP协议以及URDMA网卡分别进行吞吐性能测试。该测试通过网络损伤仪模拟广域网的丢包率和RTT时延，其中包含了多种丢包率、RTT时延情况。同时该测试使用网络测试仪按需增加测试背景流量，模拟网络带宽竞争场景。

3.2 测试结果

为公平对比，3种协议都测试单流下的吞吐性能。如图



▲图6 新型重传机制



▲图7　测试组网拓扑

8所示，随着时延（对应数据传输距离）和丢包率的增大，标准RoCEv2协议和TCP类协议吞吐急剧下降，尤其是丢包对两者的吞吐影响很大，而URDMA协议的吞吐相对稳定。

1）TCP协议的性能随时延和丢包率的变化如图8（a）。TCP在极低时延和0丢包下吞吐性能为47.6 Gbit/s。这证实其在数据中心网络中可以保持较高性能，性能瓶颈在于主机CPU和内存的配置。

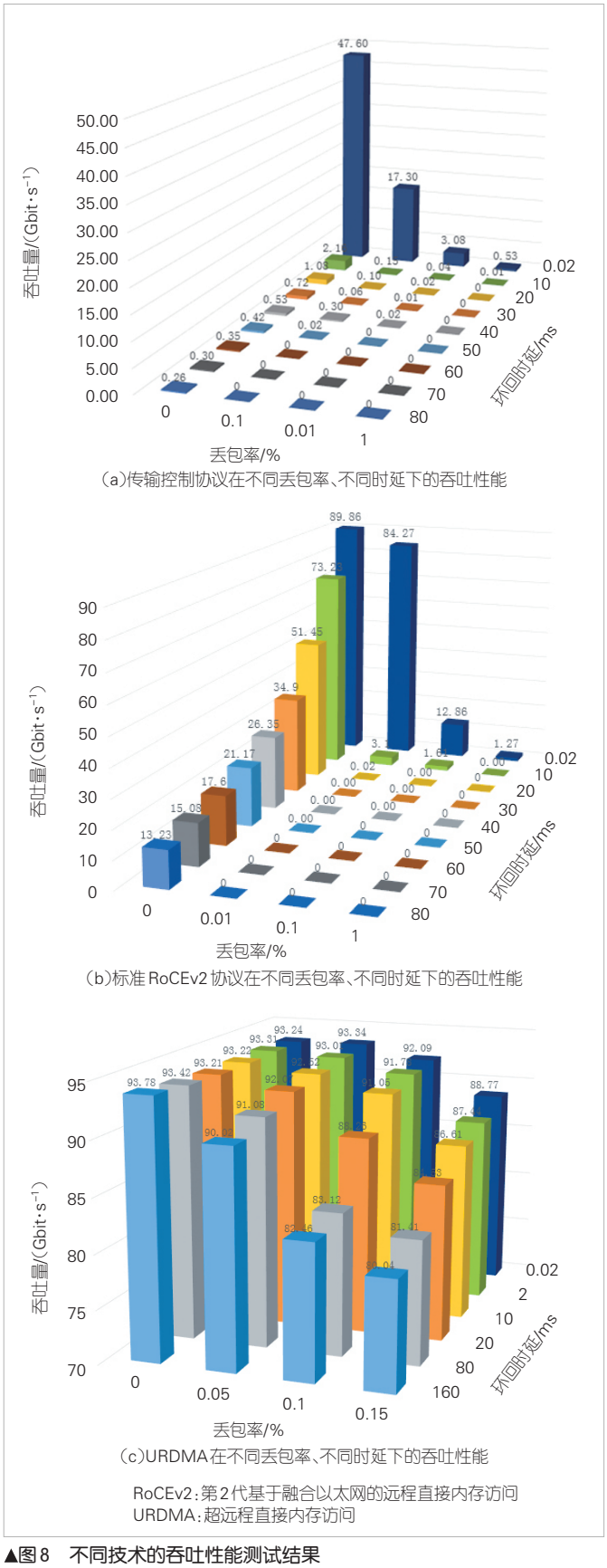
2）标准RoCEv2的性能如图8（b），其在极低时延和0丢包下也有较高性能（吞吐量达到89.86 Gbit/s）；在0丢包但时延变大的情况下，吞吐量会缓慢降低，在80 ms时衰减到13.23 Gbit/s。但如果增加丢包，RoCEv2的性能会急剧衰减，直至不可用。这表明标准RoCEv2对丢包非常敏感，只能在无损网络下使用。

3）URDMA协议性能如图8（c），其性能虽然也会随着时延和丢包变化，但衰减幅度较小。在RTT为20 ms、丢包率为0.1%的网络条件下，TCP类协议吞吐仅为0.02 Gbit/s，标准RoCEv2性能衰减到几乎为0，URDMA协议吞吐性能为88.26 Gbit/s。在RTT为80 ms时，TCP和RoCEv2协议吞吐都降为0，已不可用，URDMA协议吞吐性能为83.12 Gbit/s，仍然保持较高的性能。

4 结束语

RDMA技术已广泛应用于高性能存储、分布式训练等数据中心网络，其高吞吐、低时延性能已得到证实。RoCEv2协议凭借其良好的兼容性逐渐成为高性能网络的主流技术，但其在广域网的应用还在探索中，尚无成熟的商用方案。

本文中，我们从广域网高吞吐数据快递需求出发，分析了当前长距高吞吐协议与技术机制的不足，提出数据快递广域抗损高吞吐技术体系，并研发了URDMA算力低损耗网卡。实际测试结果表明：在广域有损网络环境下，基于广域抗损高吞吐技术的URDMA网卡吞吐性能相较于标准RoCEv2协议和TCP协议有上百倍的提升，解决了数据快递、



▲图8　不同技术的吞吐性能测试结果

跨智算集群分布式AI训练等为典型应用场景海量数据广域高效传输瓶颈问题，提升了数据流通效率。

参考文献

[1] FAST获批项目 [EB/OL]. [2025-10-09]. https://fast.bao.ac.cn/cms/category/approved_projects/

[2] 中国科学院计算技术研究所, 鄢贵海. 专用数据处理器(DPU)技术白皮书 [R]. 2021

[3] JACOBSON V, BRADEN R T, BORMAN D. TCP extensions for high performance [EB/OL]. [2025-10-08]. https://www.semanticscholar. org/paper/TCP-Extensions-for-High-Performance-Jacobson-Braden/a5fc067bca0ee49e047fffd89fc8cd2686f3be21

[4] Jacobson V. Modified TCP congestion avoidance algorithm [EB/OL]. [2024-10-09]. https://www. semanticscholar. org/paper/Modified-TCP-Congestion-Control-Algorithm-for-in-Roy/68507f828ac07a7051150785ebbcd3cfa7e3bbbe

[5] HA S, RHEE I, XU L S. CUBIC: A new TCP-friendly high-speed TCP variant [J]. ACM SIGOPS operating systems review, 2008, 42(5): 64 – 74. DOI: 10.1145/1400097.1400105

[6] TAN L S, YUAN C, ZUKERMAN M. FAST TCP: fairness and queuing issues [J]. IEEE communications letters, 2005, 9(8): 762 – 764. DOI: 10.1109/lcomm.2005.1496608

[7] BRAKMO L S, PETERSON L L. TCP Vegas: end to end congestion avoidance on a global Internet [J]. IEEE journal on selected areas in communications, 2006, 13(8): 1465 – 1480. DOI: 10.1109/49.464716

[8] CARDWELL N, CHENG Y, GUNN C S, et al. BBR: Congestion-based congestion control [J]. Communications of the ACM, 2017, 60(2): 58 – 66. DOI: 10.1145/3009824

[9] RICHARD S, FENNER B, RUDOFF A M. UNIX网络编程卷1：套接字联网API [M]. 北京: 人民邮电出版社, 2009

[10] 李博杰. 基于可编程网卡的高性能数据中心系统 [D]. 合肥: 中国科学技术大学, 2019

[11] 任宏. 关于TOE技术的发展及概况的研究 [J]. 红外, 2005, 26(3): 19-26. DOI: 10.3969/j.issn.1672-8785.2005.03.005

[12] 英特尔亚太研发有限公司. Linux开源网络全栈详解: 从DPDK到OpenFlow [M]. 北京: 电子工业出版社, 2019

[13] InfiniBand℠ Trade Association. InfiniBand™ architecture specification release 1.4 [EB/OL]. [2024-10-03]. https://www.infinibandta.org/tag/infiniband-architecture-specification/

[14] InfiniBand Trade Association. Supplement to infiniBand architecure specification volume 1 release 1.2.2 annex A 16 [EB/OL]. [2024-10-03]. https://www.infinibandta.org/tag/infiniband-architecture-specification

[15] Intel. Understanding iWARP [EB/OL]. [2024-10-05]. https://www.intel.com/content/dam/support/us/en/documents/network/sb/understanding_iwarp_final.pdf

[16] InfiniBand Trade Association. Supplement to infiniBand architecure specification volume 1 release 1.2.2 annex A 17 [EB/OL]. [2024-10-05]. https://www.infinibandta.org/tag/infiniband-architecture-specification

[17] ROCA V, BEGEN A. RFC8680 forward error correction (FEC) framework extension to sliding window codes [EB/OL]. [2024-10-06]. https://www.rfc-editor.org/rfc/rfc8680.html

[18] LUBY M, VICISANO L. RFC3695 compact forward error correction (FEC) schemes [EB/OL]. [2025-10-06]. https://www.rfc-editor.org/rfc/rfc3695.html

作 者 简 介

段晓东，中国移动通信有限公司研究院副院长、“新世纪百千万人才工程”国家级人选，教授级高级工程师；长期从事下一代互联网、算力网络、5G网络架构、6G网络架构、SDN/NFV等技术研究工作。

陆璐，中国移动通信有限公司研究院基础网络技术研究所副所长、中国通信标准化协会TC5核心网组组长；长期从事算网一体，以及移动核心网策略、演进、标准和技术研究工作，主要涉及未来网络架构、智能管道、边缘计算、算力网络等领域。

孙滔，中国移动集团首席专家，正高级工程师，中国科学技术协会第十届全国委员会委员；长期从事移动通信网架构、IP新技术研究和标准化工作。

李志强，中国移动通信有限公司研究院基础网络技术研究所高级工程师；长期从事未来IP网络演进、标准和技术研究工作，涉及下一代IP网络、算力网络、云网融合、SDN/NFV、5G核心网等。

杨红伟，中国移动通信有限公司研究院基础网络技术研究所研究员，高级工程师；长期从事下一代IP网络的技术和应用研究工作。

杜宗鹏，中国移动通信有限公司研究院基础网络技术研究所研究员，高级工程师；研究方向为算力网络、未来IP网络、确定性网络等；已发表论文10余篇。

一种存储高效的IPv6路由查找方法

A Memory-Efficient IPv6 Route Lookup Approach

姜东虹/JIANG Donghong[1,2]，郑子豪/ZHENG Zihao[1,2]，李彦彪/LI Yanbiao[1,2]

（1. 中国科学院计算机网络信息中心，中国 北京 100083；
2. 中国科学院大学，中国 北京 100049）
(1. Computer Network Information Center, Chinese Academy of Sciences, Beijing 100083, China;
2. University of Chinese Academy of Sciences, Beijing 100049, China)

DOI：10.12142/ZTETJ.202406006
网络出版地址：http://kns.cnki.net/kcms/detail/34.1228.tn.20250108.1637.004.html
网络出版日期：2025-01-09
收稿日期：2024-10-15

摘要：由于IPv6前缀比IPv4更长，如何在保证查找性能的同时提高存储效率成为一个关键挑战。现有基于Trie、哈希和三态内容可寻址存储器（TCAM）的路由查找方法在存储效率和查找性能上均存在一定局限性。提出了一种基于前缀拆分模型的集合查找方法（SetSearch），旨在实现高效的IPv6路由查找与存储。SetSearch通过前缀拆分模型实现高效的片内存储，采用路由二维映射表，避免了叶推导致的IP前段爆炸问题，从而显著降低片外存储开销。此外，SetSearch还通过基于路由二维映射表的集合查找方法，将片外访存次数减少到最多一次。基于5个真实IPv6骨干路由器转发信息表（FIB）数据集的实验评估结果表明，SetSearch在片内存储、片外存储和片外访存次数等方面展现了优异的综合性能。

关键词：IPv6；路由查找；转发信息表；高效存储

Abstract: Given that IPv6 prefixes are longer than IPv4, enhancing storage efficiency without compromising lookup performance has become a critical challenge. Existing route lookup methods—such as Trie-based structures, hashing, and ternary content addressable memory (TCAM)—have limitations in storage efficiency or lookup speed. To address these challenges, this paper proposes SetSearch, a method based on the prefix-split model, to achieve efficient IPv6 route lookup and storage. SetSearch enhances on-chip storage through prefix splitting and uses a two-dimensional routing table to prevent the explosion of IP prefixes caused by leaf-pushing, significantly reducing off-chip storage demands. Furthermore, SetSearch minimizes off-chip memory accesses to a maximum of one by utilizing a set search strategy based on the two-dimensional routing table. Experimental evaluations using five real-world IPv6 backbone router forwarding information bases (FIBs) datasets demonstrate that SetSearch offers superior performance across metrics such as on-chip storage, off-chip storage, and off-chip memory access efficiency.

Keywords: IPv6; route lookup; forwarding information base; memory-efficient

引用格式：姜东虹，郑子豪，李彦彪．一种存储高效的IPv6路由查找方法[J]. 中兴通讯技术，2024, 30(6): 31-38. DOI: 10.12142/ZTETJ.202406006

Citation: JIANG D H, ZHENG Z H, LI Y B. A memory-efficient IPv6 route lookup approach [J]. ZTE technology journal, 2024, 30(6): 31-38. DOI: 10.12142/ZTETJ.202406006

路由查找是路由器、三层交换机等网络设备的核心功能，直接影响网络中数据包的转发性能。其主要任务是根据每个到达数据包的目的IP地址，在转发信息表（FIB）中找到最匹配的IP前缀，并依据与该前缀相关联的“下一跳”信息来转发数据包。随着无类别域间路由（CIDR）[1]技术的广泛应用，路由查找从简单的精确匹配问题演变为更复杂的最长前缀匹配（LPM）问题，成为网络设备中资源消耗最为显著的功能之一。

由于IPv4地址资源的枯竭，全球正在加速部署下一代互联网协议IPv6。然而，IPv6 FIB的高效存储面临着严峻挑战。一方面，由于IPv6前缀更长，相较于IPv4 FIB，IPv6 FIB在相同规模下需要消耗更多存储资源；另一方面，当前骨干路由器中的IPv6 FIB前缀数量已接近22万条[2]，并且仍呈指数增长。因此，研究一种存储高效的

基金项目：国家自然科学基金项目（62072430）

IPv6路由查找方法对于全球平稳过渡至下一代IPv6网络具有重要的现实意义。

当前的路由查找方法主要分为3类：1）基于三态内容可寻址存储器（TCAM）的方法[3-4]。该类方法的性能较高，因为TCAM能够在一个时钟周期内完成所有IP前缀的并行匹配。然而，这类方法存在功耗高且存储容量受限的缺点。另两类是基于算法的方法。2）基于哈希的方法[5-6]。哈希方法具有较小的存储开销，但哈希冲突可能导致最差情况下路由查找的访存次数无法确定。现有方法通常依赖多次哈希，这在硬件路由器中实现成本较高，因此基于多次哈希的方法更多用于软件路由器。3）基于特里树（Trie）的方法[7-8]。基于Trie的方法是目前业界主流的路由查找方法。通常，这类方法会基于芯片内存储器的流水线结构，即将Trie的不同层映射到流水线的不同流水级，不同流水级拥有独立的存储器，以实现高速流水线式并行查找。然而，尽管业界已经提出了多种Trie压缩技术[9]，但在应对更长的IPv6前缀时，基于Trie的IPv6路由查找方法的存储效率仍然较低。

为提高FIB的存储效率，一种基于前缀拆分的新型路由查找模型被提出[10]。该模型通过分析骨干路由器FIB中IP前缀间存在大量相似性，并通过一种前缀拆分与合并的方法来捕获这种相似性。具体来说，该模型通过选择一个拆分位置P，将长度小于等于*P*的IP前缀划分为FIB1，长度大于*P*的前缀划分为FIB2；随后，FIB2中的每条IP前缀以P为界被拆分为前半部分（称为IP前段）和后半部分（称为新IP前缀）；最终，相同的新IP前缀被合并形成FIB3，与每条合并后的新IP前缀关联的是一个IP前段集合。当前基于前缀拆分模型的路由查找方法（SplitTrie）[10]将FIB1中的前缀和FIB3中的新前缀分别映射到片内基于SRAM的流水线中，而FIB3中与新前缀关联的IP前段集合则映射到片外基于DRAM的大存储介质中。查找过程中，系统分别查找FIB1和FIB3，并取两者的最优值。由于FIB3中合并了大量的新IP前缀，片内存储开销显著降低。然而，当前基于前缀拆分模型的路由查找方法仍面临两大关键挑战，限制了其实际应用：一是，基于非叶推[11]的前缀拆分模型会导致单次路由查找的片外访存次数过多，严重影响查找性能；二是，基于叶推的前缀拆分模型会导致片外IP前段集合条目爆炸，以致片外存储开销巨大。

基于前缀拆分的路由查找模型[10]，我们提出了一种存储高效的IPv6路由查找新方法（SetSearch）。SetSearch在保留前缀拆分模型带来的片内存储高效优势的同时，还兼顾了查找性能和片外存储效率。

1 背景及相关工作介绍

本节首先以单比特特里树（Uni-bit Trie）为例，介绍当前业界主流的基于Trie的流水线化路由查找方法；接着，介绍前缀拆分模型并分析其优势；最后，结合基于非叶推和基于叶推的两个具体路由查找实例，分析当前基于前缀拆分模型的路由查找方法的局限性。

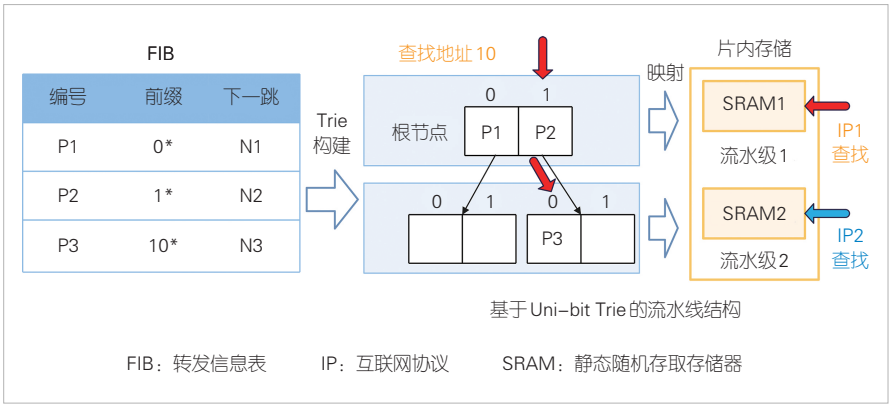
1.1 基于Uni-bit Trie的流水线式路由查找

基于Trie的路由查找方法通过将FIB组织成Trie结构，其中FIB中的每条IP前缀对应Trie中的一个实节点，且每条IP前缀由Trie根节点到其对应节点的路径表示。图1展示了如何将包含3条IP前缀的FIB构建成Uni-bit Trie结构，其中Trie节点由左右兄弟结构表示，即每个节点同时保存其兄弟节点的信息。Uni-bit Trie的查找过程从根节点开始，每次消耗IP地址的一位比特。如果该位为0，则跳转到当前节点的左孩子节点；若为1，则跳转到当前节点的右孩子节点。图1还展示了目的IP地址10的查找过程。

基于Trie的流水线式路由查找将Trie的每一层映射到流水线结构的不同流水级，每个流水级拥有独立的存储资源，因此可以实现所有流水级的并行查找。图1还展示了IP1在流水级1上查找Trie的第一层和IP2在流水级2上查找Trie的第二层在同一周期内并行执行的示例。

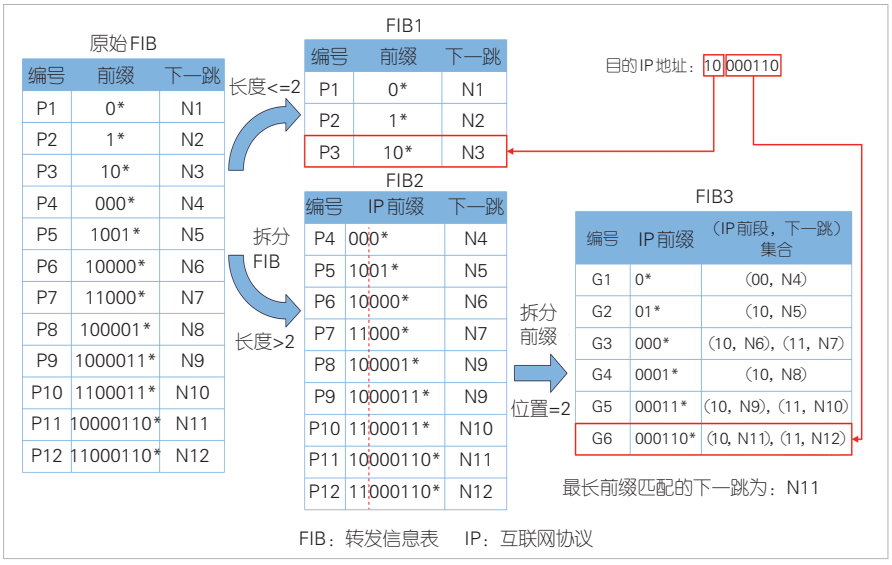
1.2 基于前缀拆分的路由查找模型及其优势分析

基于前缀拆分的路由查找模型（下文简称为前缀拆分模型）首先会选择一个拆分位置P，并依据选定的拆分位置将原始FIB拆分为FIB1和FIB2。其中，前缀长度小于等于*P*的



▲图1 基于Uni-bit Trie的流水线查找结构及IP查找示例

IP前缀被划分到FIB1中，长度大于*P*的IP前缀则划分到FIB2中。随后，以拆分位置P为界，将FIB2中的前缀进一步拆分为前半部分（IP前段）和后半部分（新IP前缀）。之后对相同的新IP前缀进行合并，形成FIB3。每条合并后的新IP前缀关联了一个IP前段集合。图2展示了对一个原始FIB进行前缀拆分的示例。在此示例中，拆分位置P选定为2，因此将长度小于等于2的0*、1*和10* 3条前缀划分到FIB1中，其余前缀划分到FIB2中。然后，依据拆分位置对FIB2中的前缀进行拆分与合并，形成FIB3。以P9和P10为例，这两条前缀的长度均大于2，因此被划分到FIB2中。拆分后P9和P10具有相同的后半部分，合并为同一条新IP前缀G5。G5关联的集合包含原P9、P10各自的IP前段及对应的下一跳信息。作为对比，FIB3中的IP前缀数目相比FIB2减少了3条。图2还展示了基于FIB1和FIB3的路由查找示例。

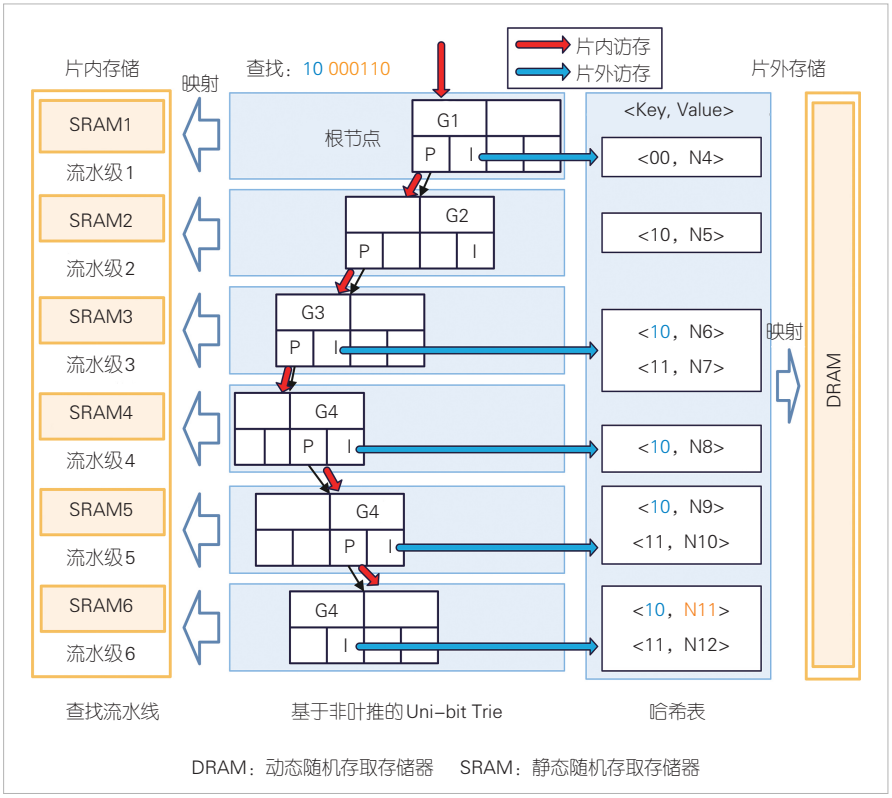


▲图2　前缀拆分过程及IP查找示例

前缀拆分模型在存储开销方面具有显著优势，原因在于经过合并后的FIB3中新IP前缀的数目相比FIB2大幅减少，能显著降低基于FIB3构建的Trie的存储开销。

1.3 当前基于前缀拆分模型的路由查找方法及其局限性

当前基于前缀拆分模型的路由查找方法SplitTrie[10]将FIB1和FIB3中的前缀分别构建为两棵不同的Trie，并分别映射到基于片内存储器（SRAM等）的流水线中。而FIB3中与新前缀关联的IP前段集合则映射到片外大容量存储介质中（DRAM等）。IP查找过程中，分别查找FIB1和FIB3，并取两者的最优值。图3展示了基于1.2节例子中的FIB3所构建的数据结构及其存储映射关系。



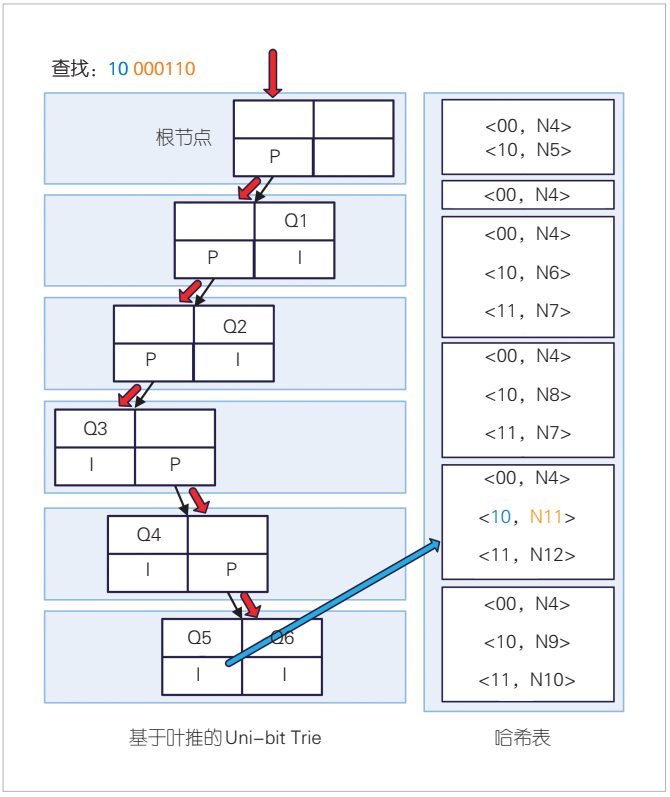
▲图3　基于非叶推的SplitTrie及IP查找示例

在IP查找过程中，SplitTrie针对FIB1的查找过程与1.1节中的例子相同。而针对FIB3进行查找时，SplitTrie首先依据拆分模型选定的拆分位置P，将IP地址拆分为前后两段。然后，使用后段在FIB3所构建的片内Trie中进行查找，每经过一个实节点（即表示匹配到一条IP前缀），则使用IP地址的前段在该IP前缀所对应的片外IP前段集合中进行查找（SplitTrie[10]使用了哈希表进行查找）。图3展示了IP地址10000110的查找过程。在该查找过程中，依次经过了G1、G3、G4、G5和G6 5个实节点，并进行了6次片外IP前段集合的查找，最终匹配到G6节点，查找结果为下一跳N11。

然而，SplitTrie将FIB3的IP前段集合（哈希表）存储在片外，因此在一次IP查找过程中，可能会进行多次片外访存。由于片外访存延迟较高，在实际应用中，单次IP查找通常只允许最多一次片外访存。过多的片外访存将显著降低IP查找性能。

为减少片外访存次数，可以引入经典的叶推技术[11]，即将所有中间实节点推送至叶子节点，使得片外访存仅发生在叶子节点，从而实现单次IP查找最多一次片外访存。图4展示了在图3示例的基础上引入叶推技术后的数据结构和相同IP地址的查找示例。在这一查找过程中，只经过叶子节点Q5一个实节点，并进行了仅一次片外访存。然而，基于叶推的SplitTrie进行叶推时，需要将FIB3中新IP前缀对应的IP前段集合一并推送至叶子节点，这会导致IP前段的大量复制，进而显著增加片外存储开销。例如，图4中叶推后的IP前段总数目由叶推前的9个增加到叶推后的15个。

综上所述，基于前缀拆分模型的非叶推SplitTrie虽然显著降低了片内存储开销，但多次片外访存严重影响了查找性能。基于叶推的SplitTrie虽然将单次IP查找的片外访存次数优化到最多一次，但叶推引发的IP前段爆炸问题会导致片外存储开销巨大。因此，当前基于前缀拆分模型的路由查找方法仍然面临严峻挑战，限制了其实际应用。



▲图4 基于叶推的SplitTrie及IP查找示例

2 基于前缀拆分模型的交集路由查找方法

为应对IPv6 FIB在存储开销方面所面临的挑战，本文中我们提出了一种存储高效的IPv6路由查找方法SetSearch。首先，SetSearch基于前缀拆分模型，在片内存储效率上表现出色；其次，利用基于路由二维映射表的集合查找方法，SetSearch在无需叶推的情况下即可实现单次IP查找仅需一次片外访存；最后，SetSearch在实现单次片外访存的同时还避免了叶推所带来的IP前段爆炸问题，相比于基于叶推的SplitTrie方法大幅降低了片外存储开销。

2.1 路由二维映射表

SetSearch是一种基于前缀拆分模型的方法。针对由前缀拆分模型生成的FIB1和FIB3，SetSearch采用不同的查找方法。首先，对于FIB1的查找，SetSearch同样使用基于片内存储器（SRAM等）的流水线式查找方法；对于FIB3的查找，SetSearch则使用了一种基于交集查找的新方法。由于针对FIB1的查找在1.1节已有详细描述，后文重点介绍针对FIB3的交集查找方法。

交集查找方法的第一步是将前缀拆分模型生成的FIB3拆解为IP前缀映射表和IP前段映射表。首先，提取FIB3中所有的新IP前缀及其对应的编号，构成IP前缀映射表。接下来，为FIB3中的每种IP前段找出所有包含该IP前段的IP前缀，并提取它们的编号以及对应的下一跳信息，生成(编号,下一跳)二元组集合。例如，在图5中的FIB3中，包含IP前段11的IP前缀包含000*、00011*和000110*，这些IP前缀的编号和

FIB3

编号	IP前缀	(IP前段，下一跳)
G1	0*	(00，N4)
G2	01*	(10，N5)
G3	000*	(10，N6)，(11，N7)
G4	0001*	(10，N8)
G5	00011*	(10，N9)，(11，N10)
G6	000110*	(10，N11)，(11，N12)

拆解

IP前缀映射表

IP前缀	编号	IP前缀	编号
0*	G1	0001*	G4
01*	G2	00011*	G5
000*	G3	000110*	G6

IP前段映射表

IP前段	(编号，下一跳)集合
00	(G1，N4)
10	(G2，N5)，(G3，N6)，(G4，N8)，(G5，N9)，(G6，N11)，
11	(G3，N7)，(G5，N10)，(G6，N12)

FIB：转发信息表 IP：互联网协议

▲图5 路由二维映射表示例

关联的下一跳二元组分别为(G3,N7)、(G5,N10)和(G6,N12)。所有IP前段和其对应的(编号,下一跳)二元组集合构成IP前段映射表。图5展示了由图2中FIB3生成的路由二维映射表示例。

2.2 数据结构及其存储映射

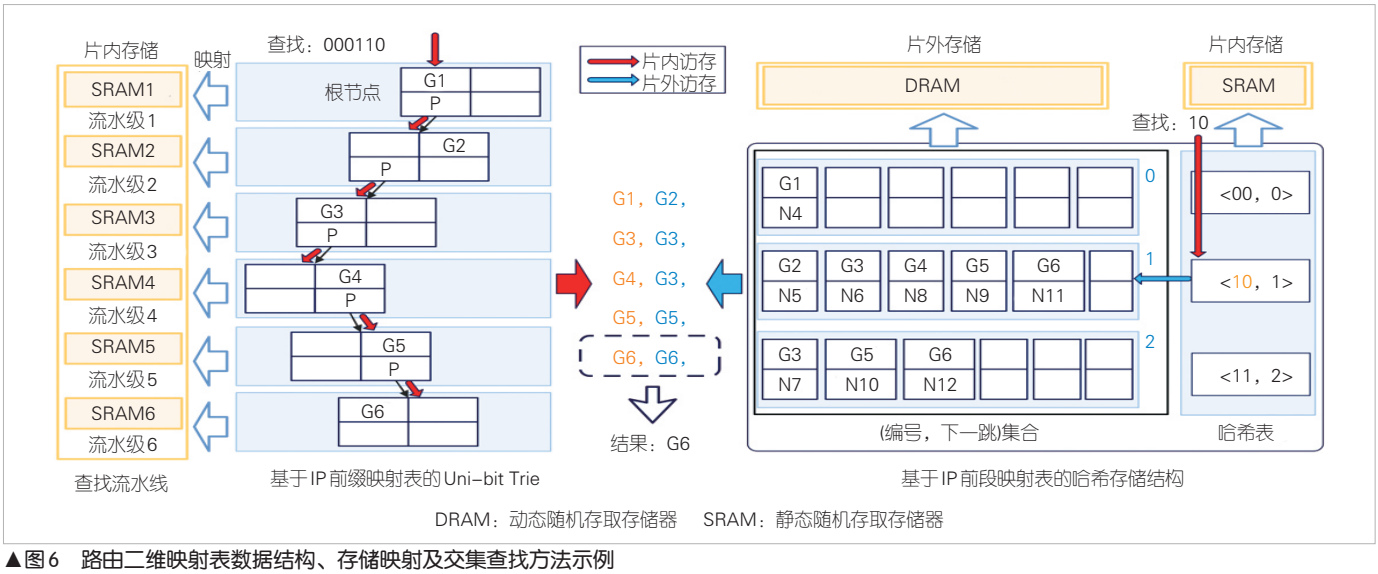
SetSearch针对FIB1的数据结构构建及其映射方式与1.1节一致。针对FIB3的交集查找方法的数据结构由两部分组成。第一部分是基于IP前缀映射表构建的非叶推Trie。该Trie的不同层被映射到流水线中的不同流水级，各流水级对应片内独立的存储器（SRAM等）。第二部分是基于IP前段映射表构建的哈希表，其中哈希表的键为IP前段，值为指向其关联的(编号,下一跳)二元组集合的索引值。哈希表被映射到片内独立存储器（SRAM等）中，而(编号,下一跳)二元组集合则被映射到片外存储器（DRAM等）中。图6展示了与图5中路由二维映射表示例相关的数据结构及其存储映射关系示例。

2.3 SetSearch查找方法

SetSearch针对FIB1的查找方法与1.1节一致。而针对FIB3的交集查找方法则分为两个阶段。第一阶段并行查找Trie和哈希表，得到两个查找结果集合；第二阶段计算这两个结果集合的交集，并选取交集中最长IP前缀对应的下一跳作为最终查找结果。如图6所示，假设待查找的IP地址为10000110，则取其后6位在映射到片内存储的Trie中进行查找，依次经过实节点G1、G3、G4、G5和G6，因此Trie的查找结果集合为{G1,G3,G4,G5,G6}。同时，取IP地址的前2位在哈希表中进行查找，命中哈希条目<10,1>，对应的(编号,下一跳)二元组集合索引值为1。进一步通过索引值从片外存储中获取查找结果集为{G2,G3,G4,G5,G6}。两个结果集的交集为{G3,G4,G5,G6}，其中G6对应的IP前缀最长，因此最终查找到的下一跳为与G6对应的下一跳N11。在这一查找过程中，系统共访问片内存储7次，其中Trie查找访问了6次，哈希表查找访问了1次；片外DRAM存储则访问1次，用于获取(编号,下一跳)二元组集合。SetSearch查找方法的完整查找过程伪代码如算法1所示。

算法1： SetSearch查找

输入：根节点 *root*, IP地址 *addr*;
输出：下一跳 *nh*.
① *SetSearch*(*root*, *nh*):
② $(addr_1, addr_2) \leftarrow$ 拆分 *addr*; // 拆分IP地址
③ $nh_1 \leftarrow root.Trie_1.lookup(addr_1)$; // 查找FIB1对应的 $Trie_1$
④ $InfoSet_1 \leftarrow root.Trie_3.lookup(addr_2)$; // 查找FIB3对应的 $Trie_3$
⑤ $index \leftarrow HashTable[addr_1]$; // 查找哈希表
⑥ $InfoSet_2 \leftarrow SetMap[index]$; // 获取二元组集合
⑦ $InfoSet \leftarrow InfoSet_1 \cap InfoSet_2$;
⑧ $nh_2 \leftarrow$ 取 *InfoSet* 中最长IP前缀的下一跳;
⑨ if $nh_1 \neq \varphi$ and $nh_2 \neq \varphi$:
⑩ $nh = nh_1.pfxlen < nh_2.pfxlen\ ?\ nh_2 : nh_1$;
⑪ else if $nh_1 = \varphi$ and $nh_2 = \varphi$:
⑫ $nh =$ 默认路由下一跳;
⑬ else:
⑭ $nh = (nh_1 == \varphi)\ ?\ nh_2 : nh_1$;
⑮ end if;
⑯ return *nh*;



▲图6　路由二维映射表数据结构、存储映射及交集查找方法示例

2.4 SetSearch查找复杂度分析

SetSearch查找方法的时间复杂度可以细分为以下几个部分。首先，IP地址拆分的时间复杂度为O(1)。其次，由于Trie查找的复杂度与地址长度线性相关，因此在$Trie_1$和$Trie_3$中进行前缀查找时，时间复杂度分别为O(|addr1|)和O(|addr2|)，其中|addr1|和|addr2|表示地址长度。哈希表的查找和SetMap的直接索引访问平均时间复杂度为O(1)。计算结果集InfoSet1和InfoSet2的交集操作，时间复杂度为O(|InfoSet1| + |InfoSet2|)。在交集InfoSet中选取最长前缀对应的下一跳，最坏情况下时间复杂度为O(|InfoSet|)。其余的条件判断和返回操作均为常数时间。因此，SetSearch方法的理论时间复杂度可表示为O(|addr1| + |addr2| + |InfoSet1| + |InfoSet2| + |InfoSet|)。然而，在实际应用中，考虑到地址长度和集合大小的限制，复杂度可以近似认为是常数级别，即O(1)。

3 实验评估与结果分析

3.1 实验设置

为了评估基于前缀拆分模型的集合查找方法（SetSearch），本文选取了以下3种方法进行对比：基于Uni-bit Trie的查找方法（Trie）、基于前缀拆分模型的Uni-bit Trie查找方法（SplitTrie）以及基于叶推的SplitTrie查找方法（SplitTrie-LP）。实验数据集来源于不同大洲的5个互联网交换中心（IXP）骨干路由器的真实IPv6 FIB数据[12]：RRC01、RRC11、RRC15、RRC19和RRC23。它们的IPv6前缀数目分别为203000、203411、210078、197051和203476。这些数据反映了2024年1月25日上午8:00的路由器状态。评估指标主要集中在片内存储开销、片外存储开销、片内访存次数以及片外访存次数。

3.2 参数选择实验

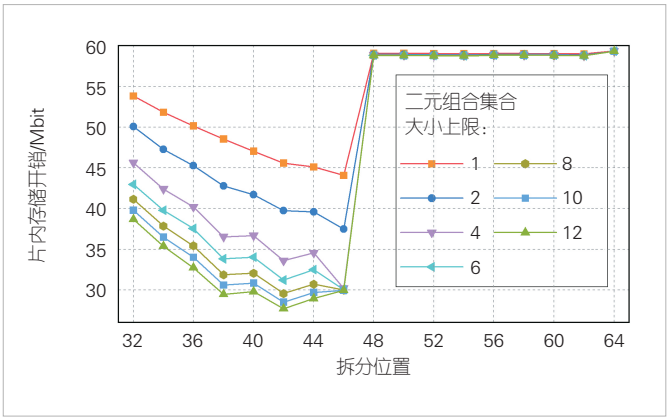
在SetSearch方法的实现过程中，片内存储开销主要受到两个参数的影响：拆分位置和与IP前段对应的(编号,下一跳)二元组集合的大小上限。其中，拆分位置决定了原始FIB中有多少IP前缀可以被拆分与合并；二元组集合大小上限决定了每个IP前段能够在片外存储的二元组数量，超出上限的二元组将导致原始IP前缀被重新划分给FIB1。

图7展示了拆分位置与二元组集合大小上限对SetSearch片内存储开销的影响。在实验中，每个二元组占用40 bit（编号和下一跳各20 bit），而单次片外访存的位宽上限设置为512 bit，因此二元组集合大小的理论最大值为12。从图7可以看出，在二元组集合大小上限小于等于6的情况下，拆分位置选择46时SetSearch的片内存储开销最小；而在二元组集合大小上限大于6的情况下，拆分位置选择42时SetSearch的片内存储开销最小。此外，图7还显示，当二元组集合大小上限为12、拆分位置为42时，SetSearch的片内存储开销在所有情况下最小。因此，后续实验中，SetSearch方法始终采用这一参数配置。

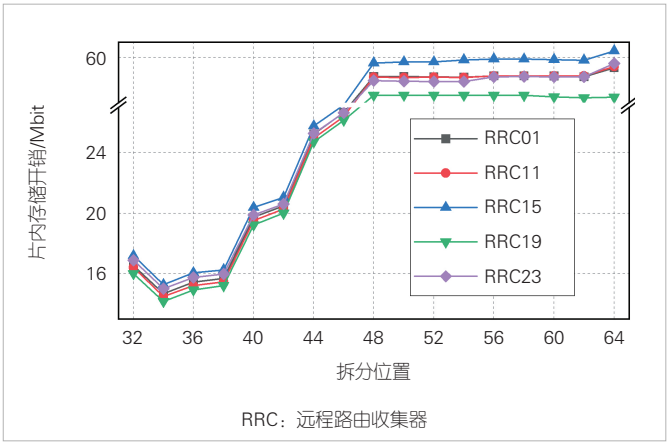
相比之下，SplitTrie方法的片内存储开销仅受拆分位置的影响。图8展示了拆分位置对SplitTrie片内存储开销的影响。从图中可以看出，在所有5个数据集中，拆分位置为34时，SplitTrie的片内存储开销最小。因此，后续实验中，SplitTrie和SplitTrie-LP方法均采用该参数设置。

3.3 对比实验

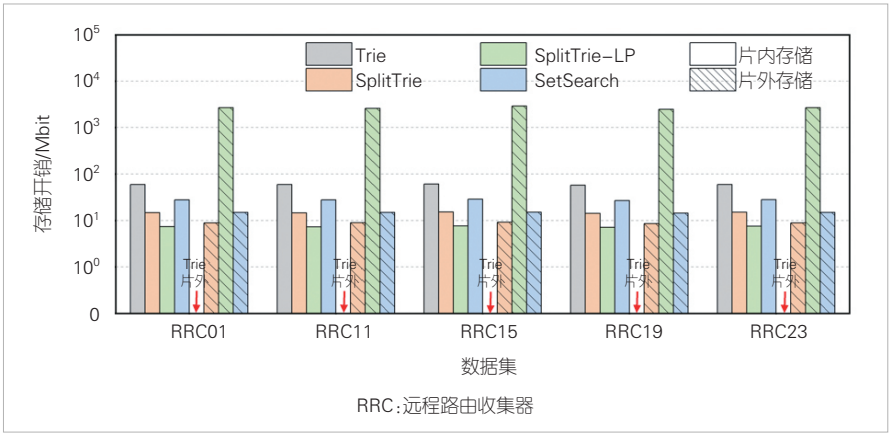
图9和图10分别展示了Trie方法、SplitTrie方法、SplitTrie-LP方法以及SetSearch方法的片内外存储开销和片内外访存次数。在存储开销方面，由于片内存储资源稀缺且昂贵，我们更关注片内存储开销；相比之下，片外存储资源更丰富且成本较低，可以适度容忍更高的片外存储开销，但



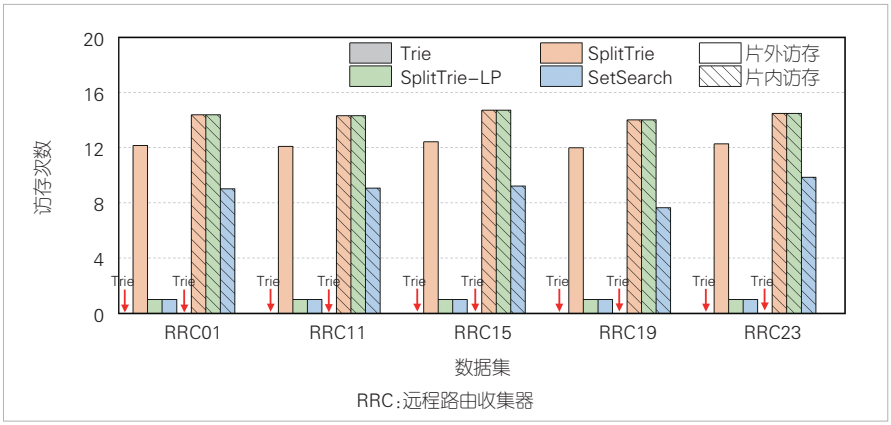
▲图7 拆分位置与二元组集合大小上限对SetSearch片内存储开销的影响



▲图8 拆分位置对SplitTrie片内存储开销的影响



▲图 9　存储开销对比



▲图 10　访存次数对比

不能过于庞大。因此，存储开销分别统计片内外的总和，即存储原始FIB的总存储开销。在访存方面，SetSearch和对比方法均采用流水线式并行查找，因此查找吞吐量仅受查找引擎主频影响，而与片内访存次数无关。然而，最大访存次数会影响查找时延。单次片内访存时延通常在纳秒级别，影响较小；但片外访存时延较高，一般的路由查找引擎最多只能容忍一次片外访存。此外，由于基于拆分模型的方法只关心FIB3的优化，因此本文只统计查找FIB3相关前缀的平均片内外访存次数。

SetSearch方法的平均片内存储开销为27.8 Mbit，平均片外存储开销为14.8 Mbit，平均片内访存次数为9次，平均片外访存次数为1次。从对比来看，虽然SetSearch方法在片内存储开销、片外存储开销以及片外访存次数等单个指标上并非最佳，但其综合表现最优，且各项性能均衡。具体来说，与Trie方法相比，SetSearch的片内存储开销平均下降了53.2%；与SplitTrie方法相比，SetSearch的片外访存次数下降了91.8%；与SplitTrie-LP方法相比，SetSearch的片外存储开销下降了99.4%。

尽管当前的SetSearch方法在综合表现方面具有较优性能，但在实现层面仍然面临一个关键挑战：如何降低IP前缀映射表和IP前段映射表查找结果的交集操作的开销。该问题的优化方向主要有两个：一是降低IP前缀映射表所构建的Trie深度，二是减小二元组集合的大小上限。

对于优化方向一，由于IPv6前缀主要集中于长度32和48之间，且3.2节中实验表明选择拆分位置为42较优，因此一种可行的方法是将FIB3中前缀长度范围限制为43至48，从而将IP前缀映射表所构建的Trie最大深度限制为8。对于优化方向二，一种可行的方法是通过选择关键比特位，将二元组集合分散到不同存储区域，从而进一步减少二元组结果集合。针对上述优化，我们将在未来的研究工作中继续探索。

4 结束语

针对IPv6路由查找中FIB存储开销大的问题，我们提出了一种基于前缀拆分模型的集合查找方法（SetSearch）。该方法基于前缀拆分模型，通过路由二维映射表的构建和集合查找机制，在不引入叶推操作的情况下，有效降低了片内和片外的存储开销，并实现了单次路由查找仅需一次片外访存。实验结果表明，与现有方法相比，SetSearch展现出了更为均衡的综合性能，在片内存储开销、片外存储开销和片外访存次数方面的综合表现最优。

参考文献

[1] FULLER V, LI T. Classless inter-domain routing (CIDR): the internet address assignment and aggregation plan [EB/OL]. [2024-10-25]. https://tools.ietf.org/html/rfc4632
[2] BGP. AS131072 IPv6 BGP table data [EB/OL]. [2024-10-25]. https://bgp.potaroo.net/v6/as2.0/index.html, 2024-09-11
[3] ZANE F, NARLIKAR G, BASU A. CoolCAMs: power-efficient TCAMs for forwarding engines[C]//IEEE INFOCOM 2003. Twenty-second Annual Joint Conference of the IEEE Computer and Communications Societies (IEEE Cat. No. 03CH37428). IEEE, 2003, 1: 42-52
[4] HE P, ZHANG W Y, GUAN H T, et al. Partial order theory for fast TCAM updates [J]. IEEE/ACM transactions on networking, 2018, 26(1): 217-230. DOI: 10.1109/TNET.2017.2776565

[5] WALDVOGEL M, VARGHESE G, TURNER J, et al. Scalable high speed IP routing lookups [C]//Proceedings of the ACM SIGCOMM '97 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication. ACM, 1997: 25–36. DOI: 10.1145/263105.263136
[6] JIANG D H, LI Y B, CHEN Y X, et al. Heuristic binary search: adaptive and fast IPv6 route lookup with incremental prefix updates [C]//Proceedings of IEEE/ACM Transactions on Networking. IEEE, Dec. 2024: 1–16. DOI: 10.1109/TNET.2024.3504244
[7] EATHERTON W, VARGHESE G, DITTIA Z. Tree bitmap: hardware/software IP lookups with incremental updates [J]. ACM SIGCOMM computer communication review, 2004, 34(2): 97–122. DOI: 10.1145/997150.997160
[8] YANG T, XIE G G, LI Y B, et al. Guarantee IP lookup performance with FIB explosion [C]//Proceedings of the 2014 ACM conference on SIGCOMM. ACM, 2014: 39–50. DOI: 10.1145/2619239.2626297
[9] ASAI H, OHARA Y. Poptrie: a compressed trie with population count for fast and scalable software IP routing table lookup [J]. ACM SIGCOMM computer communication review, 2015, 45(4): 57–70. DOI: 10.1145/2829988.2787474
[10] LI Y B, ZHANG D F, HUANG K, et al. A memory-efficient parallel routing lookup model with fast updates [J]. Computer communications, 2014, 38: 60–71. DOI: 10.1016/j.comcom.2013.10.005
[11] BIRMAN K P, SRINIVASAN V, VARGHESE G. Fast address lookups using controlled prefix expansion [J]. ACM transactions on computer systems, 1999, 17(1): 1–40. DOI: 10.1145/296502.296503
[12] RIS Docs. RIPE NCC route collectors [EB/OL]. [2024–10–25]. https://ris.ripe.net/docs/route-collectors/#bgp-timer-settings

作 者 简 介

姜东虹，中国科学院大学在读博士研究生；主要研究领域为IP查找算法、高性能路由器数据平面；发表论文3篇。

郑子豪，中国科学院大学在读硕士研究生；主要研究领域为云网络的资源管理、路由转发；发表论文1篇。

李彦彪，中国科学院计算机网络信息中心研究员；主要研究领域为高效路由转发、互联网基础资源安全与卫星互联网；主持国家重点研发计划项目、国家自然科学基金面上项目等科研项目10余项；曾获中国电子学会技术发明奖一等奖；发表论文40余篇。

智算中心网络技术发展与应用

Evolution and Applications of Network Technology in Intelligent Computing Center

段威/DUAN Wei[1,2]，李和松/LI Hesong[1,2]，周昆/ZHOU Kun[1]
（1. 中兴通讯股份有限公司，中国 深圳 518057；
2. 移动网络和移动多媒体技术全国重点实验室，中国 深圳 518055）
(1. ZTE Corporation, Shenzhen 518057, China;
2. State Key Laboratory of Mobile Network and Mobile Multimedia Technology, Shenzhen 518055, China)

DOI：10.12142/ZTETJ.202406007
网络出版地址：http://kns.cnki.net/kcms/detail/34.1228.TN.20250109.1000.006.html
网络出版日期：2025-01-09
收稿日期：2024-10-16

摘要：从应用子层、网卡子层、网络子层以及管控子层构成的完整技术栈出发，介绍了智算中心网络的关键技术。在分析智算中心网络发展趋势的基础上，介绍了中兴通讯在坚持核心自研的原则下，在芯片、产品和组网方案等方面开展的一系列创新。认为面向人工智能（AI）场景优化将成为智算中心网络发展的关键因素，行业必须在基础芯片、设备形态、网络架构、网络协议以及应用生态等方面做出更多努力，进一步推进算侧、端侧和网络侧关键技术的融合发展。

关键词：智算中心网络；大模型；以太网

Abstract: The key technologies of the intelligent computing center network are introduced from four aspects: application sublayer, network card sublayer, network sublayer, and control sublayer. ZTE Corporation has carried out a series of innovations in chip, product, and networking solutions while adhering to the principle of autonomous research and development. It is believed that artificial intelligence (AI) scenario optimization will become a key factor in the development of intelligent computing center networks, and the industry must make more efforts in the basic chip, device form, network architecture, network protocols, and application ecology to further promote the integration and development of key technologies on the computing side, end side, and network side.

Keywords: intelligent computing center network; large language model; ethernet

引用格式：段威, 李和松, 周昆. 智算中心网络技术发展与应用 [J]. 中兴通讯技术, 2024, 30(6): 39-47. DOI: 10.12142/ZTETJ.202406007
Citation: DUAN W, LI S H, ZHOU K. Evolution and applications of network technology in intelligent computing center [J]. ZTE technology journal, 2024, 30(6): 39-47. DOI: 10.12142/ZTETJ.202406007

以大语言模型（LLM）为基础的生成式人工智能（AI）技术因其良好的通用性与泛化能力，正在快速引领新一轮的科技革命和社会产业的变革。当前越来越多的科技公司竞相推出千亿、万亿参数规模的通用和垂直大模型[1]，例如OpenAI的GPT系列、Meta的LLaMA系列、百度的文心一言大模型、阿里巴巴的千问大模型等。当模型的参数规模超过数百亿后，AI大模型的语言理解能力、逻辑推理能力以及问题分析等能力迅速提升。以GPT3.5为例，参数规模达1 750亿，GPT4的参数规模更是达到了1.8万亿，参数规模的增加提升了大模型处理复杂问题的能力。训练这类超大参数规模的大模型给智能计算基础设施带来了前所未有的挑战，通常需要几千甚至数万张图形处理器（GPU）加速卡并行协同工作[2]。基于数据并行、流水线并行和张量并行等多种并行技术的分布式并行计算是实现AI大模型训练的关键手段。业界普遍认为支撑万卡以上规模的GPU集群高效运行的瓶颈在于网络互联能力[3]。如何提供一种无损、超高带宽、超低延迟、超高稳定性且可高度扩展的网络互联方案逐渐成为行业研究和关注的焦点。

1 智算中心网络特点和性能需求

大量的理论分析和工程实践表明，为确保昂贵的算力集群资源的高效利用，面向GPU互联的智算中心网络在组网架构、流量模型和性能指标3个方面和传统的数据中心网络有所不同。因此，我们需要推动智算中心网络性能的跨越式发展。

1.1 智算中心网络的组网架构特点

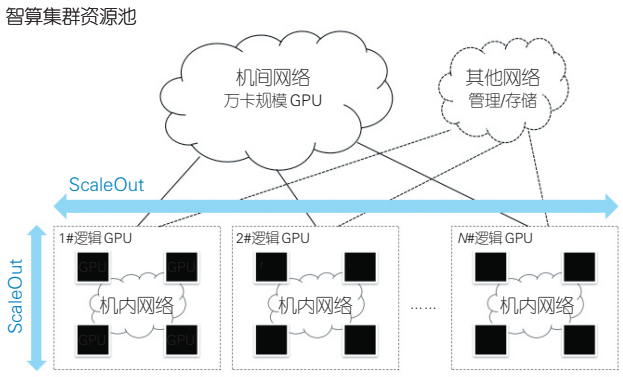
如图1和表1所示，与传统数据中心不同的是，除了传

统的管理、存储等通用网络以外，智算集群通常存在机内网络和机间网络。其中，机内网络实现的是有限数量（通常为几个到几百个）GPU之间超高速互联，通常称为ScaleUp网络。其主要目的是将多个GPU组织成一个逻辑上的超级GPU，以满足大模型并行切分的需求。此类网络通常需要数百GB的超高通信带宽、几百纳秒的低延迟以及一定的缓存一致性要求。典型的例子是英伟达通过NVlink/NVswitch构建256卡的SuperPOD。机间网络的目的是构建超大规模的GPU集群（如万张GPU卡集群规模），通常被称为ScaleOut网络。该类网络通常需要单端口100 Gbit/s以上的无收敛组网技术，由网卡和交换机组成。典型的此类网络为InfiniBand（IB，一种用于高性能计算的计算机网络通信标准）网络或者基于融合以太网的远程直接内存访问（RoCE）网络。

在实际业务运行的过程中，机内网络和机间网络由一套高性能通信库来实现灵活的资源调配。由此可见，智算中心是网络一个高度复杂的软硬件技术栈，而不是一个单点技术或者产品。

1.2 智算中心网络的流量模型特点

智算集群的业务特性决定了其在网络侧呈现独特的流量模型：低熵、大象流、同步效应。其中，低熵表示同一时刻网络中的活跃数据流（通常由五元组决定）数量较少，通常为千级，远远低于通算集群动辄数万甚至数十万的流数，这将导致传统基于流的负载均衡技术在智算中心网络中不再适用；大模型的参数规模和GPU的超高并发处理能力导致智算中心网络中大象流占据主导地位，根据中兴通讯的现网统计，单条流的峰值带宽即可达到上百GB；同步效应指的是AI训练和推理任务充斥着大量的集合通信，同一任务内的不同流之间存在明显的同步需求，导致网络的长尾延迟对业务性能影响被显著放大[5]。



GPU：图形处理器

▲图1　智算中心网络的组网架构

智算中心网络这种独特的流量模型会直接影响到现有技术的有效性，这也是推动智算中心网络技术创新的原动力。

1.3 智算中心网络的性能指标需求

结合AI大模型的业务特点和流量模型特征，中兴通讯从组网规模、带宽、延迟、可靠性以及可管理性等几个维度对智算中心网络的性能指标需求开展定性和定量分析。

1）性能指标定性分析

LLM训练、LLM推理、智能推荐等常见AI业务场景对网络的定性需求如表2所示，其中带宽指的不仅仅是网络物理带宽，还指基于物理网络能达到的业务有效吞吐；时延指的是业务转发延迟，包括网络静态转发延迟以及由于排队、丢包重传导致的动态延迟；尾延迟是评价网络延迟性能稳定性的重要指标，在AI场景中对业务性能影响巨大。

2）性能指标定量分析

从网络的视角来看，我们通常希望同样的产品和解决方案能够尽可能覆盖多样化的业务场景。综合考虑业务场景的需求，中兴通讯提炼出5个方面的定量指标，同时也分析出

▼表2 典型人工智能业务场景网络需求

业务场景	集群规模	带宽	时延	尾延迟	稳定性	可管理性
LLM训练	超大	超高	一般	较高	超高	超高
LLM推理	较小	一般	较高	较高	高	高
智能推荐	较小	较高	一般	一般	高	高

LLM：大语言模型

▼表1 智算网络的分类和定位

网络类型	网络定位	网络特点	典型技术
机内网络	ScaleUp网络，将多个GPU组成一个逻辑上的超级GPU	超高带宽，GPU点到点通信带宽百GB以上；扩展性受限，通常为几百个GPU以内；超低延迟；缓存一致性等	NVlink/Nvswitch
机间网络	ScaleOut网络，实现集群的横向扩展，组成超大的智算集群	单端口超过100 Gbit/s无收敛组网，由网卡和交换机组成，通常要求支持RDMA	Infiniband、Ethernet
其他网络	管理/存储/用户南北向流量	有收敛比的通用数据中心网络	Ethernet

GPU：图形处理器　RDMA：远程直接内存访问

了当前网络实际能力与这些指标的差距，并给出了优化方向，具体如表3所示。

2 智算中心网络的关键技术和发展趋势

智算中心网络是一个高度复杂的软硬件技术栈。本章中，我们将从关键技术和行业趋势两个方面展开具体分析。

2.1 智算中心网络的关键技术

如图2所示，从技术架构层面分析，智算中心网络主要包含应用、网卡、网络和管控4个技术子层。为了满足AI业务的性能需求，各个子层需要引入如下一些核心技术。

1）应用子层核心技术

在智算中心网络的应用场景中，网络能力通常被封装成高性能通信库供上层应用逻辑调用，典型的通信库如lib-fabric、消息传递接口（MPI）、英伟达集合通信库（NCCL）等。通信库通常需要根据具体的物理网络能力进行适配才能达到最佳性能，如拓扑发现、拓扑亲和性、负载均衡路径发现、集合通信加速等，这是智算中心网络不可或缺的一个环节。

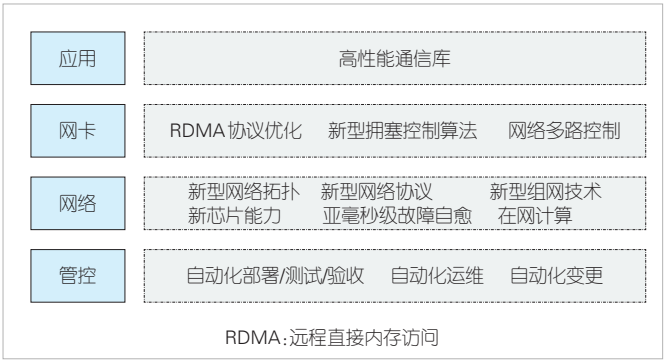
2）网卡子层核心技术

网卡是智算中心网络流量的起点和终点，很多关键功能都必须依托网卡来构建。网卡子层主要包括如下关键技术。

（1）RDMA协议及其优化

远程直接内存访问（RDMA）是网络实现高吞吐性能的基础。然而，在大规模组网环境下，传统RDMA协议容易成为性能瓶颈[6]。为了在大规模集群网络中实现高性能的RDMA传输，需要对网卡进行一系列优化，包括协议设计和传输模式。这些优化旨在解决传统RDMA协议在扩展性方面的不足。例如北京邮电大学提出的FedRDMA[7]，将模型分块之后再使用RDMA传输，优化了跨数据中心联邦学习场景中的通信效率；南京大学提出了一种可扩展的RDMA传输方法[8]，通过有效共享连接来提高性能。

（2）新型拥塞控制算法



▲图2　智算中心网络的关键技术

拥塞控制是高性能网络要解决的重要技术问题。传统拥塞控制算法如数据中心量化拥塞通知（DCQCN）[9]在扩展性、流公平性以及时延性能方面存在瓶颈，需要新型拥塞控制算法的创新。例如，阿里巴巴提出的高精度拥塞控制（HPCC）[10]，通过利用In-band网络遥测，实现了高精度的拥塞控制，显著提升了网络的吞吐量并降低了延迟性能。Google提出的Swift[11]拥塞控制算法，依靠基于延迟的反馈机制，进一步降低了延迟，并提高了吞吐量。Poseidon[12]通过使用带内遥测（INT）携带拥塞信息，只响应网络中瓶颈点的拥塞信号，从而达到了高效拥塞控制的目的，在快速收敛的同时实现最大公平性。新加坡国立大学与诺基亚合作提出的LinkGuardian[13]则采用链路本地重传机制，有效解决了数据中心网络中的链路丢包问题，减少了丢包率，并维持了较低的延迟和高吞吐量。中国科学技术大学提出的低时延高精度拥塞控制（LHCC）[14]使用带外遥测技术，能够快速通知网络状态，使发送端在一个往返时延（RTT）内感知到拥塞情况，并根据路径上所有队列的状态调整发送速率。这种机制使得LHCC在网络中存在多个瓶颈时依然能提供更精确的拥塞控制。中国科学院计算技术研究所提出的FlexPass[15]是一种基于Credit的传输协议，结合加权公平队列和双控制循环，在保持低延迟的同时实现高吞吐量和零丢包。这些新型拥塞控制算法通过引入高精度监控和智能反馈机制，进一步推动了数据中心网络的性能提升，特别是在减少延迟、提高

▼表3　智算网络定量需求分析

智算网络需求	需求描述	当前数据中心网络能力	优化方向
万卡集群规模	按照GPU节点数计算，集群规模按照场景分级，私有云资源千卡级，AI工厂/公有云万卡级	传统RDMA集群数小于1 000	智能网卡、物理网络
TB互联带宽	单台8卡服务器接入带宽1.6T+，有效吞吐>90%	接入10G/25G，有效吞吐<60%	物理网络、负载均衡
极致网络性能	0丢包，微秒级时延，微秒级低抖动	时延亚毫秒级，拥塞情况下秒级，基础无损	网络拓扑、拥塞控制
超高网络稳定性	亚毫秒级故障恢复，性能一致性	50 ms级检测，秒级/分钟级收敛，性能不稳定	管控、网络设备
多维网络自动化	部署、验收、运维、变更自动化	能力参差不齐，整体欠缺	管控、网络设备

AI：人工智能　GPU：图形处理器　RDMA：远程直接内存访问

吞吐量和减少丢包方面取得了显著进展。

（3）网络多路径控制

网络多路径是智算中心网络吞吐性能的基础，也是在故障场景下保障通信性能的重要基础手段。路径发现、探测以及报文喷洒是网络多路径控制的主要功能。新加坡国立大学提出的ConWeave[16]通过网络内的重路由和数据包重新排序，为RDMA流量提供了有效的负载均衡，显著改善了平均和高百分位的流完成时间。湖南大学的RDMA轻量级快速报文重排机制（LEFT）[17]则利用双状态共享位图方案减少了内存消耗，并通过快速和慢速路径的数据包重新排序降低了延迟，即使在多路径RTT差异较大的情况下，仍能保持高吞吐量。复旦大学的基于主机的流片微调（HF2T）[18]进一步优化了RDMA负载均衡，通过在主机端延迟少量数据包，延长数据包间的时间间隔，增加flowlet的生成机会，从而提升flowlet负载均衡的效果。清华大学基于向量协议的RDMA拥塞感知负载均衡（CAVER）[19]则采用了基于向量协议的拥塞感知负载均衡方法，通过利用确认字符（ACK）包在网络中传播拥塞信息，实时为源服务器机柜顶交换机（ToR）提供最不拥塞的路径。在不改变现有硬件的前提下，CAVER通过对ACK包携带的路径拥塞信息进行传播，帮助源交换机快速找到最优路径。

3）网络子层核心技术

网络层是由交换机组成的高速互联网络。为了更好地满足智算业务的实际需求，网络本身需要引入一系列的技术创新，包括新型网络拓扑、新型网络协议、新型组网技术、新芯片能力、亚毫秒级故障自愈以及在网计算等。这些技术涉及网络芯片、网络设备、网络协议等多个方面，也是当前技术创新最活跃的领域之一。

（1）新型网络拓扑

智算中心的网络拓扑有很多技术选择，包括CLOS、FatTree、BiGraph[20]、Dragonfly[21]以及Torus[22]拓扑，每一种拓扑都有其各自的优点和适用场景。新型网络拓扑及其配套的路由协议是智算中心网络需要关注的重要内容。

（2）新型网络协议

网络组网拓扑和性能需求的变化必然导致新型网络协议的出现。为了能在新场景中实现最佳的网络路由、高效的负载均衡，需要设计新的网络协议。这也是目前业界研究的重点，典型的协议如胖树网络路由协议（Rift）[23]、自适应路由[24]等。

（3）新型组网技术

针对智算中心网络的组网需求，目前业界提出了很多新型组网技术，包括分布式分散式机框（DDC）[25]、全调度以太网（GSE）[26]等。他们大多通过一种封闭式Fabric设计以期达到最优的网络性能。该领域也是当前创新最为活跃的领域。

（4）新芯片能力

网络芯片的优化空间广泛，除了传统的可编程能力外，Buffer管理和时延优化是两个重要的方面。目前明确采用共享缓存架构的厂商有思科、博通以及英伟达，且作为其主要的架构亮点，低时延是需要持续优化改进的方向，其本质是采用更低层次的Cut-Through逻辑。

（5）亚毫秒级故障自愈

为了提高AI算力集群的资源利用率，网络需要具备能够达到亚毫秒级的故障自愈能力。由于智算中心网络通常都有多条等价路径可用，收敛性能的瓶颈则在于故障感知能力。当前基于echo-reply机制的故障检测基本都是50 ms级，需要芯片层引入一些快速故障感知甚至故障预测的能力。

（6）在网计算

在网计算是当前智算中心网络研究的热点，也是英伟达网络互联方案的核心。通过网络交换机支持集合通信卸载加速，可以获得计算任务提速、网络拥塞缓解的双重收益。根据英伟达给出的测试数据，其在网计算的实现方案可扩展的分层聚合与规约协议（SHARP）[27]可带来2倍以上的计算任务加速，其集群规模越大，收益越明显。在未来机内互联和机间互联的场景中，在网计算带来的整体收益将日益明显，这也将成为未来智算中心网络的关键技术需求之一。

4）管控子层核心技术

对于上万节点规模的智算集群而言，网络的部署、测试、验收、运维和变更等需要依托管控平台构建完善的自动化能力。这种管控平台的管理对象不再限制为网络本身，而是逐渐覆盖到应用、网卡和网络等端到端的通信链条，这也是与传统网络的管控系统最大的区别。

各子层所涉及的关键技术的基本原理以及与需求的对应关系如表4所示。

2.2 智算中心网络的发展趋势

AI大模型让智算产业空前繁荣，智算中心网络也进入了技术创新的高速发展周期。通过对当前行业和技术的洞察分析，中兴通讯认为当前智算中心网络呈现如下几个明显的发展趋势。

趋势1：RDMA成为智算中心网络高性能协议的主流技术

传统数据中心业务通常对吞吐和延迟有较高的容忍度，简单易用的传输控制协议/互联网协议（TCP/IP）即可满足

▼表4 智算网络的关键技术分析

场景	需求	关键技术	功能主体	技术原理
智算网络	万卡级集群规模	新型网络拓扑	交换机	面向智算的特定场景（如LLM训练），在交换机能力一定的情况下，通过新型网络拓扑提升集群规模和网络效率
		RDMA协议优化	网卡	针对传统RDMA协议的不足，开展协议优化，以支撑集群向万卡以上规模扩展（如从RC传输模式转向RD传输模式等）
	TB级互联带宽	TB级接入	交换机	基于高容量交换机芯片支持100G/400G/800G高密度组网
		网络多路径能力	网卡、交换机	通过端网协同完成流量的多路径传输，提升网络吞吐性能
	极致网络性能	新型拥塞控制算法	网卡	通过拥塞控制算法的创新，确保业务能在集群规模扩展的前提下低延迟和低抖动
		新芯片能力	交换机	在交换芯片层面针对AI场景引入新的特性，增强故障倒换和拥塞感知能力
		新网络协议	交换机	针对新型网络拓扑和新的拥塞故障场景，研究对应的网络协议
		新型组网技术	交换机	研究新型组网技术如DSF、全调度以太网GSE等
		在网计算	网卡、交换机	网卡和交换机协同设计，实现集合通信的卸载加速功能
	超高稳定性	故障快速检测	交换机	交换芯片提升故障感知能力
		性能一致性	交换机	研究网络硬隔离和软隔离的技术方案

AI：人工智能　GSE：全调度以太网　RC：可靠连接模式　RDMA：远程直接内存访问
DSF：分布式调度网络　LLM：大语言模型　RD：可靠数据报模式

业务的承载需求。智算中心网络本质上一个高性能网络的需求。传统TCP/IP协议已经无法满足业务对高吞吐、低延迟、零算力损耗的要求，RDMA逐渐成为智算中心网络高性能协议的主流技术。智算中心网络未来发展的核心目标之一是支持超大规模的RDMA高性能集群组网。

趋势2：采用以太网来构建智算中心网络成为更广泛的行业共识

以太网已成为智算中心网络最主流的支撑技术。过去10年里，以太网凭借完善的标准体系、成熟的产业链、快速演进的接口速率、灵活的向后兼容能力等优点，成为当前数据中心网络采用的事实标准。目前以太网接口已经发展到800G/1.6T，交换芯片容量从100G迅速提升到51.2T/102.4T，容量增长近100倍，单比特功耗下降90%以上。将以太网延伸到智算中心网络各个场景已经成为行业共识，面向AI场景优化的新型以太网技术正处于快速发展期。在这种趋势的引领下，国际上成立了超级以太网联盟（UEC），中国也出现了全调度以太网和高通量以太网等一系列创新成果。相信以太网将成为未来智算中心网络最基础、最主流的支撑技术。

趋势3：场景融合成为智算中心网络技术的创新方向

随着AI集群组网规模的持续增长，多场景融合将是未来智算业务对网络的内在需求。场景融合具体体现在如下几个方面。

1）总线和网络的融合

当前AI集群网络由负责垂直扩展的总线型互联和负责横向扩展的网络互联构成两个独立异构的互联域。当集群规模进一步扩展到十万节点甚至百万节点的规模时，这种组网方式的成本和可维护性将难以为继。随着互联技术的快速演进，总线和网络的统一承载成为可能，网络总线化和总线网络化的趋势将成为智算中心网络技术创新的主要方向。

2）多业务的融合承载

智算业务的发展日新月异，混合专家模型（MoE）和多模态技术会带来不同的网络侧需求，流量模型、带宽需求、时延指标等呈现多样化的特征，需要网络具备“一网多用”的综合承载能力。多业务综合承载不仅是未来智算中心网络面临的挑战，也是需要通过技术创新解决的核心问题之一。

3）光电技术融合组网

从英伟达最新的NVL72/576互联架构[28]来看，电互联和光互联融合组网是构建高能效比超级算力集群的关键。光电混合组网技术在以谷歌为代表的下一代数据中心中大规模部署落地[29]。大量的实践数据表明，光电混合组网在构建高性能、低能耗、低成本的AI网络中效果明显[30]，是下一代智算中心网络架构演进的趋势。

场景融合必然带来网络技术的跨域融合，这也是是智算中心网络下一阶段创新的热点方向。

趋势4：新型大容量网络芯片成为智算中心网络发展的基石

大容量网络芯片是AI数据中心高速互联的基础载体，主要包括数据处理器（DPU）网卡芯片和网络交换芯片。DPU网卡芯片是流量的网络入口，是RDMA、拥塞控制、各种加速引擎的功能载体。网络交换芯片数据中心交换机最核心的部件，决定着网络的组网规模和整体性能。当前国际最先进的DPU网卡芯片已经具备400G/800G接口速率，51.2T的交换芯片已经规模化落地，可以支撑3万张新一代GPU集

群。随着智算业务对高速互联的需求持续攀升，网络芯片正处于一个高速发展的阶段，呈现出如下明显的发展趋势。

1）容量持续增长，单比特功耗持续降低

过去10年，以太网交换芯片容量从百G迅速提升到51.2T，容量增长近100倍，单比特功耗下降90%以上。在AI的驱动下，未来交换网络芯片容量将迅速突破100T，单比特功耗进一步降低。与此同时，400G/800G DPU网卡需求也将迎来井喷。

2）面向AI场景优化成为网络芯片发展的基本要求

在过去两年里，新一代网络芯片引入面向AI场景优化的新特性成为行业主旋律。这也将是未来5~10年网络芯片更新迭代的主要推动力。典型技术包括超低延迟、故障预测、智能流分析引擎、基于容器/包的负载均衡、在网计算等。

面向AI场景优化的新型大容量网络芯片是智算中心网络发展的基石，需要在高速接口、交换架构等基础技术方面持续创新突破，是自主可控需要重点关注的核心技术之一。

3 智算中心网络的典型应用场景和解决方案

智算中心网络在应用场景上存在一定的多样性，行业解决方案也呈现百家争鸣的繁荣景象。中兴通讯做了大量的实践和理论分析，本章节我们将对应用场景和行业解决方案分别进行概述。

3.1 智算中心网络的典型应用场景

按照业务类型分，除了传统通算场景已经覆盖的虚拟私有云（VPC）网络、存储网络和管理网络以外，智算中心网络聚焦在AI加速器（包括GPU和DPU等）之间的高速互联网络，包括通常意义上的机内ScaleUP网络和机间ScaleOut网络两部分。中兴通讯认为，应用场景需要分类分级，准确把握应用场景的需求是匹配最佳网络方案的基础，具体如下。

1）不同业务类型对网络的需求差异较大

如表2所示，典型的业务类型包括LLM训练、LLM推理、智能推荐等，不同的业务对网络性能的要求不同[31]。具体而言，LLM训练是典型的带宽密集型业务，延迟容忍度较高，流量的轨道效应明显；LLM推理的Decode是延迟敏感性业务，延迟通常决定着推理集群的资源利用率；智能推荐通常要求较高的带宽和延迟性能。这种业务类型的差异会直接影响网络方案的选择。

2）用户的目标集群规模分级明显

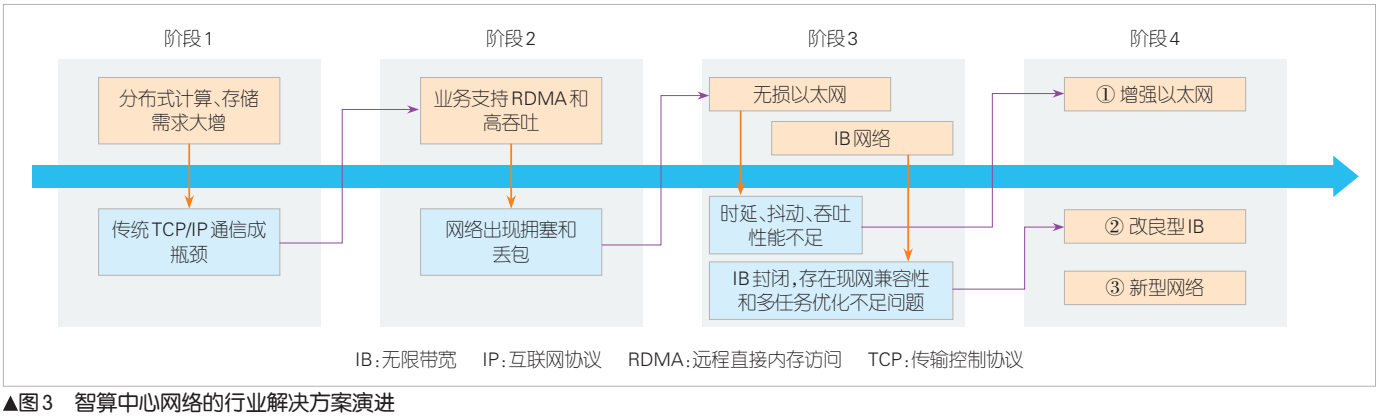
当前行业掀起了万卡以上甚至十万卡、百万卡规模智算集群研究的热潮，而真正需要万卡以上规模集群的应用通常是LLM基础训练。据统计，未来真正有能力投资建设万卡以上资源池来从事基础大模型训练的企业不到10%。这意味着90%以上的企业建设智算资源池是用来进行大模型的精调和推理，而满足这类需求的集群规模通常只需要千卡级以下规模的集群。在千卡以下的智算集群中，网络方案可以做得极其简单，过于复杂的方案会导致用户付出额外成本。中兴通讯认为，未来智算中心网络的建设将进入一段时间的冷静期，从用户的实际需求出发匹配最具性价比的方案才是网络价值的回归。

3.2 智算中心网络的行业解决方案

智算中心网络的行业解决方案有很多（如图3所示），但从技术路线来看，中兴通讯认为目前主要有增强以太网、改良型IB和新型网络3种行业解决方案。3种技术路线各具特点。

1）增强型以太网

增强型以太网的基本思路是在现有以太网技术的基础上



▲图3 智算中心网络的行业解决方案演进

针对智算场景的需求开展体系化的优化，以满足智算中心网络的业务需求。目前业界普遍认可的是采用端网协同优化以太网在拥塞控制、时延、抖动以及吞吐方面的性能。该技术路线是阿里、腾讯等头部互联网公司的一致选择，也是中兴通讯智算中心网络的主要产品路线。

2）改良型IB网络

长期以来，英伟达将IB作为其智算中心网络的主要解决方案，这给业界传递了一种不太准确的信息：IB是最适合智算中心网络的方案。事实上，在实际的智算场景测试中，以太网在带宽、组网规模等方面均不逊于IB[32]。从技术角度分析，IB网络并非是为当前新兴的AI业务量身定制的，而是传统高性能计算（HPC）市场的方案传承。英伟达宣称的大部分IB技术优势如自适应路由、SHARP在网计算等均是基于无限带宽网络贸易协会（IBTA）标准的私有化改良。随着智算中心网络的发展，IB网络也将不断演进和改良，并在相当长一段时间内维持其在智算中心网络领域的市场份额。

3）新型网络

由于智算场景对规模和性能的极致追求，行业内开展了很多新型网络技术的探索和实验。这类方案的特点是试图最大限度重用现有产业链：在网络拓扑、互联技术方面创新，追求极致性能。但这样会牺牲一定的异厂商互通性。典型的技术方案有开放计算项目（OCP）提出的分布式调度网络（DSF）方案，以及中国移动提出的GSE方案。

4 中兴通讯在智算中心网络中的技术和产品创新

基于对智算中心网络需求和行业趋势的洞察，中兴通讯在坚持核心技术自研的前提下，沿着增强以太网和新型网络两条路线开展了一些列技术和产品的创新，形成了完整的智算中心网络解决方案。

1）基于自研芯片的智算交换机产品

以全自研芯片为基础，中兴通讯已经形成了完整的智算系列交换机产品（如图4所示），包括盒式59和框式99两大系列。59盒式系列交换机设备单机容量达到12.8T，并将快速迭代到单机51.2T，性能和可编程能力达到业界先进水平。99框式系列交换机设备采用中国性能最高、国际领先的自研分布式芯片，可以提供多达576个400G端口密度，单层网络即可支持万卡以上规模集群的组网需求。

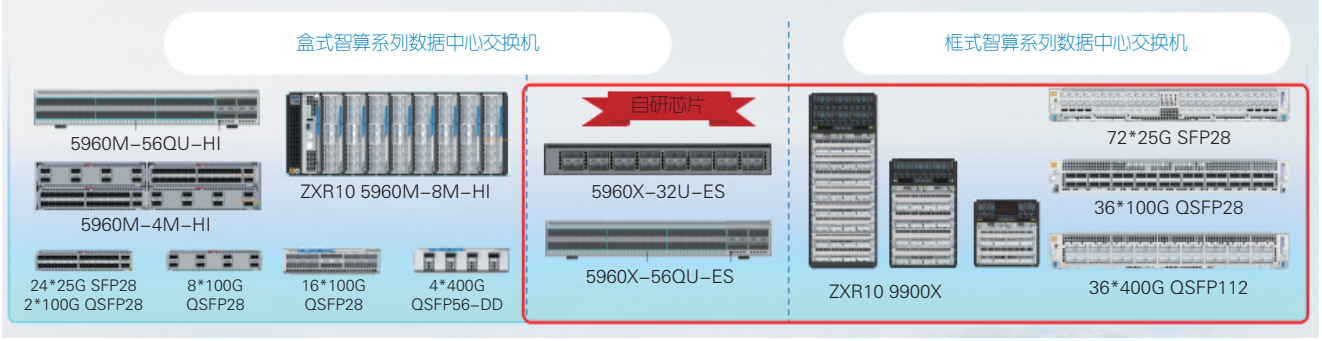
2）智算中心网络技术和方案创新

在智算中心网络的技术路线上，中兴通讯沿着增强以太网和新型网络方面开展技术和方案创新，形成了差异化的解决方案。

（1）增强以太网的技术创新实践

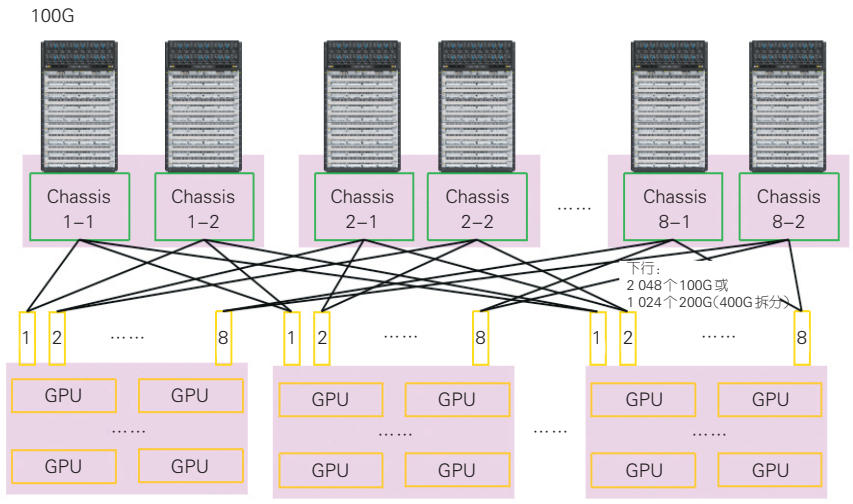
通过自研DPU产品和交换机之间的端网协同设计，中兴通讯解决了网络拥塞状态的精细化感知、网络多路径、全局负载均衡、端网统一管控等一系列工程化难题。

中兴通讯在智算网络技术上不断创新，通过轨道负载分担（ZRLB）技术，基于增加交换机入端口（Ingress Port）作为Hash Key的算法，根据业务需求把连接服务器的端口进行Group分组，并基于Group内物理口进行Hash，实现出口流量均衡，提升网络负载均衡效率；通过智能全局负载分担（iGLB）技术，实现“网络控制器+AI调度平台”协同工作，精确掌握每条业务流的带宽诉求，集中计算出每条业务流的最优转发路径，达到整网负载均衡效率最佳；通过端网协同拥塞控制（ENCC）技术，带内遥测扩展链路状态信息，端网协同实现精准拥塞控制、链路故障实时反馈、网络快速发送拥塞通知报文，从而改进流量调度；同时控制器通告整网多路径信息，端侧DPU根据流负载均衡选路，网络按规划路径转发，保证智算业务流量能够充分利用网络多路径资源，提高传输吞吐率，实现端网协同的新型精准拥塞流控。



QSFP：四通道小型可热插播接口　SFP：小型可插拔式

▲图4　中兴通讯智算系列交换机产品



▲图5　中兴通讯单层多轨双平面组网方案

通过自适应路由通告（ARN）技术，依靠芯片硬件级支持能力，对链路的拥塞、中断故障快速“检测、传递、切换”，实现端到端的路径切换时间小于1 ms。

（2）新型网络的方案创新实践

以自研分布式芯片和对应的业界最高密度框式交换机为基础，中兴通讯采用单层多轨双平面的创新组网方案（如图5所示），提升网络可靠性和网络规模，确保将网络侧的故障影响降到最低，完美支持16K A800集群的组网需求。相比传统方案，该方案网络层次极简，带宽利用率接近100%，大幅度降低光模块的互联成本。

5 结束语

智算中心网络作为支撑AI业务的重要基础设施，需要应用子层、网卡子层、网络子层以及管控子层构成的完整技术栈的方方面面一起协作、创新，以提供超大网络规模、无损、低时延、高吞吐、高可靠以及高可维能力等高性能网络的技术特性。智算中心网络在未来几年将迎来市场和技术的跨越式发展。综合考虑当前AI大模型的发展趋势和中国算力基础设施的现状，支撑百万级集群规模应为智算中心网络的基本要求。但是面向中国单点算力建设规模受限、算力碎片化严重的现状，大数据入算的新场景、跨域分布式训练、高通量数据传输等新需求陆续出现，对网络提出了更高的要求，规模上的量变将带来技术上的质变。网络在支持大模型训练的同时，还需要具备训推一体、支持多租户隔离的网络架构，从而推进大模型更广泛的应用。未来几年，面向AI场景优化将成为智算中心网络发展的主旋律，行业必须在基础芯片、设备形态、网络架构、网络协议以及应用生态等方面做出更多努力，以进一步推进算侧、端侧和网络侧关键技术的融合发展。中兴通讯相信开放繁荣的智算中心网络生态才是行业的未来，并将持续在该领域做出更多的原创性成果。

参考文献

[1] KODALI R K, PRASAD UPRETI Y, BOPPANA L. Large language models in AWS [C]//Proceedings of 1st International Conference on Robotics, Engineering, Science, and Technology (RESTCON). IEEE, 2024: 112–117. DOI: 10.1109/restcon60981.2024.10463557
[2] YELURI S. Large language models: the hardware connection [EB/OL]. [2024–10–10]. https://blog. apnic. net/2023/08/10/large–language–models–the–hardware–connection
[3] TANG Z H, SHI S H, WANG W, et al. Communication–efficient distributed deep learning: a comprehensive survey [EB/OL].(2020–03–10)[2024–10–06]. https://arxiv.org/abs/2003.06307
[4] NVIDIA. NIVIDIA spectrum–X network platform architecture [EB/OL]. [2024–10–06]. https://resources. nvidia. com/en–us–accelerated–networking–resource–library/nvidia–spectrum–x
[5] CISCO. Evolve your AI/ML network with Cisco silicon one [EB/OL]. [2024–10–06]. https://www. cisco. com/c/en/us/solutions/collateral/silicon–one/evolve–ai–ml–network–silicon–one.html
[6] ZHANG Z, LUO L, NING Q, et al. SRNIC: a scalable architecture for RDMA NICs [EB/OL]. [2024–10–03]. https://www.usenix.org/conference/nsdi23/presentation/wang–zilong
[7] ZHANG Z L, CAI D Q, ZHANG Y R, et al. FedRDMA: communication–efficient cross–silo federated LLM via chunked RDMA transmission [EB/OL]. (2024–03–01)[2024–10–08]. https://arxiv.org/abs/2403.00881
[8] TANG J, WANG X L, DAI H C. Scalable RDMA transport with efficient connection sharing [EB/OL]. [2024–10–05]. https://ieeexplore.ieee.org/document/10228968
[9] ZHU Y, ERAN H, FIRESTONE D, et al. Congestion control for large–scale RDMA deployments [EB/OL]. [2024–10–05]. https://dl.acm.org/doi/10.1145/2785956.2787484
[10] LI Y, MIAO R, L H H, et al. HPCC: high precision congestion control [EB/OL]. [2024–10–07]. https://dl. acm. org/doi/10.1145/3341302.3342085
[11] KUMAR G, DUKKIPATI N, JANG K, et al. Swift: delay is simple and effective for congestion control in the datacenter [EB/OL]. [2024–10–12]. https://dl.acm.org/doi/pdf/10.1145/3387514.3406591
[12] WANG W, MOSHREF M, LI Y, et al. Poseidon: efficient, robust, and practical datacenter CC via deployable INT [EB/OL]. [2024–10–11]. https://www.usenix.org/conference/nsdi23/presentation/wang–weitao
[13] JOSHI R, SONG C H, KHOOI X Z, et al. Masking corruption packet losses in datacenter networks with link–local retransmission [C]//Proceedings of the ACM SIGCOMM 2023 Conference. ACM, 2023: 288–304. DOI: 10.1145/

3603269.3604853
[14] YAN B, ZHAO Y, XU S, et al. LHCC: low-latency and hi-precision congestion control in RDMA datacenter networks [EB/OL]. [2024-10-06]. https://ieeexplore. ieee. org/document/10682889
[15] LIM H, KIM J, CHO I, et al. FlexPass: a case for flexible credit-based transport for datacenter networks [EB/OL]. (2023-05-08) [2024-10-07]. https://dl.acm.org/doi/10.1145/3552326.3587453
[16] SONG C H, KHOOI X Z, JOSHI R, et al. Network load balancing with In-network reordering support for RDMA [C]//Proceedings of the ACM SIGCOMM 2023 Conference. ACM, 2023: 816-831. DOI: 10.1145/3603269.3604849
[17] HUANG P, ZHANG X, CHEN Z, et al. LEFT: lightwEight and fast packet reordering for RDMA [EB/OL]. (2024-08-03) [2024-10-08]. https://dl.acm.org/doi/abs/10.1145/3663408.3663418
[18] CHEN C, YE J, GAO Y, et al. HF2T: host-based flowlet fine-tuning for RDMA load balancing [EB/OL]. [2024-10-10]. https://dl.acm.org/doi/10.1145/3663408.3663410
[19] DENG H T, YANG Y, ZHANG M, et al. CAVER: enhancing RDMA load balancing by hunting less-congested paths [EB/OL]. [2024-10-10]. https://dl.acm.org/doi/10.1145/3672202.3673729
[20] DONG J B, CAO Z, ZHANG T, et al. EFLOPS: algorithm and system co-design for a high performance distributed training platform [C]//Proceedings of IEEE International Symposium on High Performance Computer Architecture (HPCA). IEEE, 2020: 610-622. DOI: 10.1109/hpca47549.2020.00056
[21] KIM J, DALLY W J, SCOTT S, et al. Technology-driven, highly-scalable dragonfly topology [EB/OL]. [2024-10-12]. https://ieeexplore.ieee.org/document/4556717
[22] JOUPPI N P, KURIAN G, LI S, et al. TPU v4: an optically reconfigurable supercomputer for machine learning with hardware support for embeddings [EB/OL]. (2023-04-04)[2024-10-05]. https://arxiv.org/abs/2304.01433
[23] IETF. Routing in fat trees (RIFT) working group [EB/OL]. [2024-10-05]. https://datatracker.ietf.org/doc/draft-ietf-rift-rift/
[24] NVIDIA. Adaptive routing [EB/OL]. [2024-10-05]. https://docs.nvidia. com/networking-ethernet-software/cumulus-netq-48/Monitor-Operations/Monitor-Adaptive-Routing/
[25] OPEN COMPUTE PROJECT 2020. Distributed disaggregated chassis routing system [EB/OL]. [2024-04-10]. https://www.opencompute. org/documents/ufispace-dcc-routing-system-intro-for-ocp-summit-2020-1-pdf
[26] 段晓东, 程伟强, 王瑞雪, 等. 面向新型智能计算中心的全调度以太网技术 [J]. 中兴通讯技术, 2023, 29(4): 57-63. DOI: 10.12142/ZTETJ.202304011
[27] NVIDIA. SHARP: in-network scalable streaming hierarchical aggregation and reduction protocol [EB/OL]. [2024-10-12]. https://mug. mvapich. cse. ohio-state. edu/static/media/mug/presentations/20/bureddy-mug-20.pdf
[28] NIVIDA. NIVIDA GB 200 NVL72 [EB/OL]. [2024-10-13]. https://www.nvidia.com/en-us/data-center/gb200-nvl72/
[29] POUTIEVSKI L, MASHAYEKHI O, ONG J, et al. Jupiter evolving: transforming google's datacenter network via optical circuit switches and software-defined networking [EB/OL]. [2024-10-12]. https://dl.acm.org/doi/pdf/10.1145/3544216.3544265
[30] WANG W Y, GHOBADI M, SHAKERI K, et al. Rail-only: a low-cost high-performance network for training LLMs with trillion parameters [EB/OL]. [2024-10-13]. https://ieeexplore. ieee. org/document/10664412
[31] RABINOVITSJ D. Opening AI infrastructure [EB/OL]. [2024-10-13]. https://drive. google. com/file/d/1ud1JZqco2868AvmkNkrA-Axp-74PvwWx/view
[32] BROADCOM. Why ethernet reigns supreme over InfiniBand for large-scale networks [EB/OL]. [2024-12-04]. https://docs.broadcom.com/doc/Unleashing-the-Power-of-AI-ML

作 者 简 介

段威，中兴通讯股份有限公司资深研发总工，高级工程师；主要从事IP网络、智算中心网络关键技术研究；申请专利30余项。

李和松，中兴通讯股份有限公司技术规划专家，主要从事智算、光电子以及网络相关的技术研究和规划工作。

周昆，中兴通讯股份有限公司研发规划有线总工，负责AI智算、金融DC等多个业务领域的DCN方案和产品规划。

超以太网技术的现状与展望

Status and Prospect of Ultra-Ethernet Technology

厉俊男/LI Junnan，李韬/LI Tao，杨惠/YANG Hui
（国防科技大学，中国 长沙 410073）
(National University of Defense Technology, Changsha 410073, China)

DOI：10.12142/ZTETJ.202406008
网络出版地址：http://kns.cnki.net/kcms/detail/34.1228.TN.20250108.1617.002.html
网络出版日期：2025-01-08
收稿日期：2024-10-17

摘要：随着数据中心、智算中心规模的急剧增长，传统以太网技术在通信带宽和延时等方面面临巨大挑战。深入分析传统以太网的优缺点，从物理层、链路层、传输层和软件层4个方面梳理了超以太网技术，并对其中的关键技术展开详细的介绍和研究。此外，还分析了超以太网相关技术在中国的发展现状。最后，探讨了超以太网技术发展面临的机遇与挑战。

关键词：超以太网；网络协议；人工智能

Abstract: With the rapid growth of data centers and intelligent computing centers, traditional ethernet protocols face enormous challenges in terms of communication bandwidth and latency. The advantages and disadvantages of traditional ethernet are analyzed. Then the key technologies of super Ethernet are introduced from four aspects: physical layer, link layer, transmission layer, and software layer. In addition, the development status of related technologies in China is also analyzed. Finally, the opportunities and challenges faced by the development of Ultra Ethernet technology are discussed.

Keywords: ultra ethernet; network protocols; artificial intelligence

引用格式：厉俊男, 李韬, 杨惠. 超以太网技术的现状与展望 [J]. 中兴通讯技术, 2024, 30(6): 48-53. DOI: 10.12142/ZTETJ.202406008
Citation：LI J N, LI T, YANG H. Status and prospect of ultra-ethernet technology [J]. ZTE technology journal, 2024, 30(6): 48-53. DOI: 10.12142/ZTETJ.202406008

随着人工智能（AI）技术的飞速发展，AI对算力的需求呈现指数级增长的态势，这促使围绕大规模分布式计算的基础设施建设迅猛发展。在庞大的分布式系统中，网络作为连接各个节点的“神经脉络”，不仅是数据流通的管道，更是基础设施互联的粘合剂。网络将算力资源、存储资源以及各类智能应用紧密地结合在一起。然而，面对AI技术日益增长的复杂性和高性能要求，现有的网络技术如以太网和IB（InfiniBand）网络，在各自领域内表现出色，但逐渐显露出局限性。

AI技术的发展迫切需要一种更加先进、高效、灵活且成本可控的网络解决方案。这种网络需要具备更大规模的扩展能力，以应对不断增长的算力需求；需要更高的带宽，以确保数据传输的畅通无阻；需要支持多路径传输，以提高网络的可靠性和容错性；需要实现对拥塞的快速反应和智能调度，以保证数据传输的实时性和稳定性；同时，还需要充分考虑单个数据流执行度的相互依赖关系，特别是尾延迟这一关键因素，以确保AI应用的整体性能和用户体验。

基于这样的背景，我们有必要重新审视和评估现有的网络技术，积极探索和研发能够满足智能计算需求的新型网络技术。这不仅是一次对网络技术的重大挑战，更是推动AI技术持续发展的关键所在。

1 超以太网技术发展背景

超以太网传输（UET）架构主要是从物理层、链路层、传输层与软件层4个方面来改进以太网技术，既兼容现有的以太网生态，又能提升以太网的交换转发性能，从而改进存储、管理、安全结构，提升遥测能力。超以太网传输技术由业界领军企业组成的非盈利性组织——超以太网联盟（UEC）提出，其目的是优化现有以太网技术，开发高性能全栈架构，满足当前人工智能（AI）对网络性能、灵活性和成本效益的严苛需求，推动相关技术的研发、标准制定及市场推广，以引领未来网络技术的发展方向。

基金项目：国家重点基础研究发展计划项目（2010CB328200、2010CB328201）；国家高技术研究发展计划项目（2006AA01Z257）；国家自然科学基金项目（60602058、60572120）；国家科技重大专项项目（2009ZX03003-002-02）

1.1 以太网的优势与面临的挑战

以太网自1973年诞生至今，获得了巨大成功：速率从最早的10 Mbit/s发展到如今的100 Gbit/s、200 Gbit/s甚至400 Gbit/s；广泛应用于各类AI训练的大型集群中。以太网/IP协议族具有众多优势：

1）具有极好的通信生态。以太网协议已经十分成熟，拥有广泛的应用市场。支持以太网协议的包括以太网交换机、网卡、线缆、收发器、光电转换等设备厂商，以及相应的以太网管理工具厂商。以太网使用标准的网络设备和标准化的通信协议，这使得部署和维护成本较低。

2）支持高带宽互连。以太网高达每秒数百吉比特的传输速率，可为数据中心提供计算资源之间的高速互联，也可为用户提供高速的网络资源访问能力，以满足现代网络用户对速度和效率的需求。

3）具备较高的可靠性。以太网通信是一种可靠的通信技术，采用错误检测和冗余机制，可以保证关键任务或者敏感数据传输的完整性和正确性。

4）易于管理。以太网不仅管理结构相对简单，同时具有丰富的网络管理工具和配置协议，能够有效简化网络管理员的配置和网络监管，提升管理效率。

5）配套使用的IP协议也非常成熟。IP网络支持大规模的路由寻址，能够支持机架级、园区级和数据中心级网络。

以太网的众多优势造就了其在AI计算领域的广泛应用。随着AI模型对算力需求的急剧增加，网络成为互联分布式计算资源的关键，并在AI大模型训练中变得越来越重要。

大型语言模型（LLM）如GPT-3、Chinchilla和PALM，推荐系统如深度学习推荐模型（DLRM）、深度和层次化集成网络模型（DHEN），都是在数千个图形处理器（GPU）的集群上进行训练的[1]。这些大型语言模型通常采用分布式训练方式，不同计算节点间存在频繁数据交互过程，即每启动新的一轮计算需要等待所有计算节点完成上一轮计算和数据交互。不同节点间数据交互过程最后一个消息到达的时间决定了下一轮计算阶段启动的时间。因此，尾部延迟通常是AI分布式计算系统性能的关键指标。

大型模型的参数数量持续增加，上下文窗口范围持续扩大。例如，2020年GPT-3拥有1 750亿参数[2]，而最近发布的GPT-4模型已有近一万亿参数[3]，DLRM更是拥有数万亿参数[1]，并仍会继续增长。这些规模愈发庞大的AI模型需要更大的集群以提供相应的训练算力，配套更高的通信带宽以实现数据交互。与此同时，网络时延也愈发重要，如网络的延时拥塞会造成集群中昂贵计算资源的闲置。

1.2 超以太网联盟

UEC由AMD、博通、思科、英特尔、Meta和微软等10家来自芯片、通信、互联网行业的领导者于2023年牵头成立，旨在完善以太网标准，以更好地满足人工智能、机器学习和高性能计算不断增长的需求[4]。

目前，UEC发展迅速，除了牵头的10家厂商外，已有超80家知名厂商加入该联盟，包括芯片设计、计算、通信、互联网等主流厂商，如IBM、Candence、Synopsys、瞻博网络、戴尔等。中国厂商也积极加入该联盟，如中兴通讯、华为、新华三、百度等。其中，阿里巴巴加入UEC技术委员会，与Meta、AMD、博通和微软等其他12名成员，一同推进以太网核心计算的研发工作和相关标准制定工作。

UEC成立之初划分了4个工作组，分别是物理层、链路层、传输层和软件层工作组。

1）物理层工作组。该工作组制定以太网物理层规范、电气和光信号特性规范，开发应用程序接口和定义相关数据结构，以提高物理层传输性能，降低传输延迟，改善以太网物理层配置管理。当前物理层工作组主要制定100 GbE和200 GbE速率端口物理层协议（PHY）规范。目前已经确定了100 GbE介质类型、PHY支持的速率和类型，200 GbE的规范还在制定中。

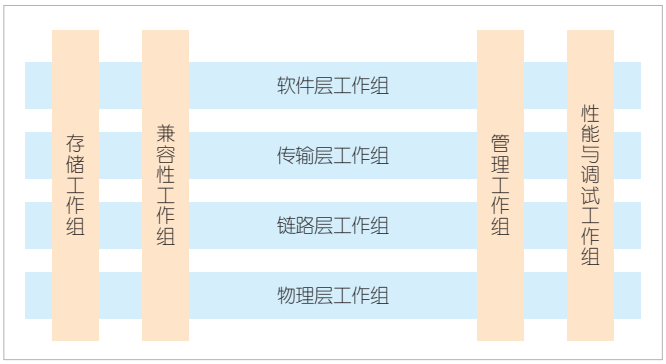
2）链路层工作组。该工作组主要研究链路层可靠性、多路径与报文喷洒策略、链路特性协商机制，以提升链路层传输的可靠性、传送效率和遥测能力。

3）传输层工作组。通过研究拥塞控制算法、安全策略、宽松的报文重排序机制，传输层工作组避免基于以太网的远程直接内存访问（RoCE）传输可能存在尾延时大的缺点，解决报文可靠传输、数据安全传送、应用程序扩展等难题。

4）软件层工作组。该工作组采用了兼容现有通信库的方法。现有通信库包括集合通信库（CCL）、信息传递接口（MPI）、共享内存（SHMEM）通信库等。它们使用libfabric作为数据平面框架的应用程序编程接口（API），这有助于上层应用的快速开发和部署。同时还定义了加速器和高速扩展接口（FEP）之间的交互方式，即各类加速器API。通过同一交换机、FEP以及聚合管理器（AM）的控制平面和数据平面接口定义，解决不同UEC供应商之间的互联互通、互操作问题。

近期，UEC又成立了存储、管理、兼容性与测试、性能与调试工作组，如图1所示，旨在完善超以太网系统应用、互操作等方面的能力。在此基础上，超以太网联盟已发布超以太网白皮书1.0版本[5]。

此外，UEC与其他开源联盟关系密切，如开放计算项目



▲图1 超以太网联盟工作组划分

联盟（OCP）、全球网络存储工业协会（SNIA）、OpenFabric联盟（OFA）、IEEE 802.3工作组等。

2 超以太网关键技术

为提升以太网传输带宽，降低尾延时，我们从物理层、链路层、传输层和软件层4个方面来研究各层协议的优化技术。图2展示了超以太网协议栈整体架构，物理层旨在提升以太网速率，支持100～200 Gbit/s；链路层则是优化传输可靠性和传输性能；传输层主要从拥塞控制、可靠性传输、数据安全、链路遥测4个方面优化设计；软件层/应用层通过拓展协议库为上层应用提供相应服务。

2.1 物理层技术

物理层除了上文提及的制定以太网物理层规范、电气和光信号特性规范，开发应用程序接口和定义相关数据结构外，还研究链路质量预测与评估的概念，并制定相关指标如误码率（UCR）、PHY平均错误时长（MTBPE）、平均误包被接受时长（MTTFPA），以更精确地预测和度量物理层链路质量。其中，误码率用于标识链路上数据报文发生错误的频率，PHY平均错误时长用于标识PHY接口的错误率，平均误包被接受时长用于标识误收错误报文的比例。

2.2 链路层技术

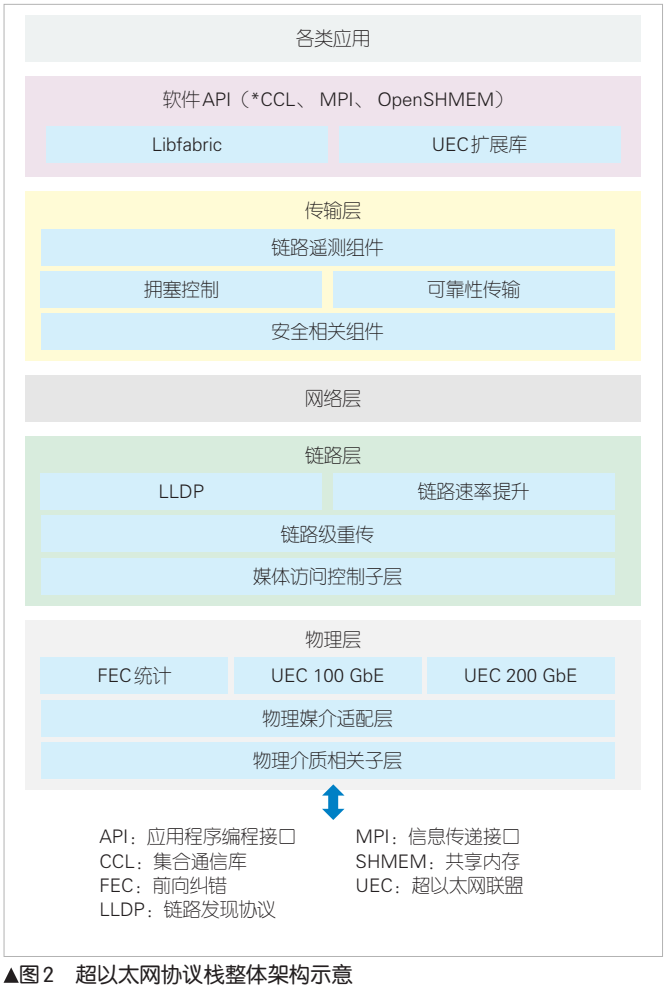
链路层工作组从可靠性、报文传输效率以及多路径与报文喷洒3个方面来提升链路层传输可靠性和传输性能。

1）链路可靠性保证机制。该机制通过在链路层的逻辑链路管理（LLC）和MAC管理之间设计和插入新的子层——链接级别重试（LLR），以构建链路层端到端错包重传。

2）报文传输效率提升策略。该策略针对智能计算大量消息有效载荷在16字节的特点，以及传统以太网短报文有效载荷比率过小的问题，采用以太网报文头压缩策略，增加帧的传送效率。为兼容现有以太网协议，报文头中设计了压缩标识信息，用于区分压缩报文和非压缩报文，从而允许两类报文可在网络中共存，而不影响原有的功能。

3）多路径与报文喷洒。传统的以太网网络基于生成树协议，确保从A到B的单一路径，以避免网络中的环路。随后出现了多路径技术，例如等价多路径（ECMP）[6]，网络尝试利用尽可能多的链路来连接通信对象。ECMP通常使用“流哈希”，它将不同五元组的流量映射到不同路径上，也可以将不同五元组的流量映射到同一条路径上。然而，这种方法可能将高吞吐量流量限制在一条路径上，当过多的流量映射到单一网络路径时，网络性能会下降，因此需要对负载均衡进行精细管理以获得最佳性能。

超以太网的基本思路是将单条流的不同报文同时分散到所有可以通往目的地的路径上，这种技术被称为“报文喷洒”，可以更加均衡地利用所有的网络路径。这种更灵活的多路径策略会引入报文频繁乱序的问题。如果仍然采用严格的报文排序要求，则会阻止乱序报文直接从网络传输到应用程序缓冲区，最终限制传输效率。



▲图2 超以太网协议栈整体架构示意

在AI工作负载中，大量GPU或者加速器之间数据需要频繁交互。这其实是一种“集合”通信，包含All-Reduce和All-to-All两种模式，其中All-Reduce通过单节点上获取所有节点信息，并执行Reduce操作；All-to-All作为全交换操作，通过All-to-All通信，可以让每个节点都获取其他节点的信息。考虑到AI应用程序只关心给定消息的最后部分何时到达目的地，集合通信快速完成的关键是节点间快速完成批量传输。针对上述两种交互方式，采用报文喷洒可有效降低数据交互的尾延时。

2.3 传输层技术

传输层工作组从可靠性传输、数据安全、拥塞控制、链路遥测4个方面展开研究。四者的关系如图3所示，拥塞控制与链路遥测密切配合，通过感知链路状态以准确、及时反馈拥塞状态；可靠传输与拥塞控制构成报文传输模块；安全模块负责数据的加解密以及密钥的分发管理。

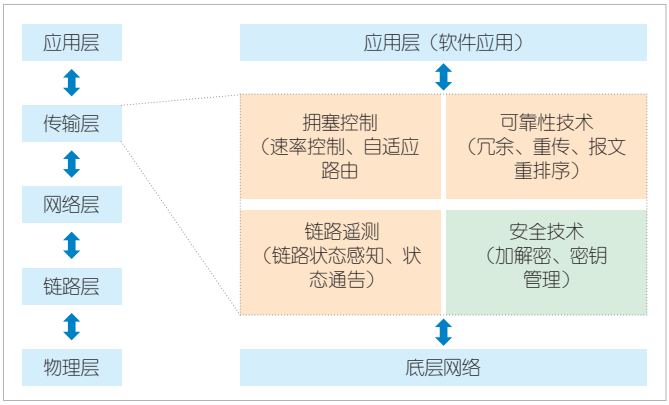
1）支持报文乱序的可靠传输技术。全链路的报文喷洒必然会引入更多的报文乱序，为此针对不同应用需求可设置3类不同的报文传输模式：

（1）可靠，有序传输（ROD）。该模式按照顺序传输报文，用于需要消息有序传输的应用。

（2）可靠，无序传输（RUD）。该模式只能向语义层传输一次报文，但可以忍受网络中的乱序传输。可靠性传输层需检测重复报文，以确保每个报文只能向语义层传送一次。

（3）不可靠，无序传输（UUD）。不可靠报文可以承载许多UET的新语义，用户无须可靠传输，通过其他方式就可保障可靠性。

2）安全传输。安全传输机制作为超以太网传输技术的重要研究内容，可针对业务的需求，以及任务对延时、吞吐率的要求，选择报文头、部分报文或者全部报文数据负荷加密和认证。



▲图3 传输层关键设计

3）拥塞控制技术。网络拥塞可能发生在3个地方：发送方到第一跳交换机的出站链路、第一跳交换机和最后一跳交换机之间的链路、最后一跳交换机到接收方的最后链路。对于AI来说，发送方出站链路上的拥塞主要可以通过发送主机上的调度算法进行控制，因为主机可以看到所有出站流量。上文中提到的多路径报文喷洒通过均匀分配所有路径上的负载，最小化了第一跳和最后一跳交换机之间的热点和拥塞。拥塞的另一种形式——“Incast”，发生在多个发送者同时向同一目标发送流量时[7]，即最后一条到接收方的链路上。Incast既可能发生在“All-Reduce”过程，也可能发生在“All-to-All”过程。

近年来，学术界和工业界对拥塞控制展开了广泛研究，提出了许多优秀的拥塞控制算法，如数据中心量化拥塞通知（DCQCN）[8]、数据中心TCP拥塞控制协议（DCTCP）[7]、简单有效的拥塞控制（SWIFT）[9]、Timely[10]等。但是上述拥塞控制算法无法同时满足为AI优化的传输协议的所有需求，这些需求包括：

（1）在高带宽、低往返时间（RTT）网络中，当链路无拥塞时，整个网络可快速达到线速，而不对路径已存在的流量造成影响，即降低已有流量的吞吐率。

（2）感知整个网络的拥塞程度，并充分利用多路径最大限度提升传输效率。

（3）公平共享最后一跳链路来避免incast、报文丢失、重传或尾部延迟。

（4）与流量特点、硬件架构解耦，无须随着流量组合的变化、计算节点的发展、链路速度的提高和网络硬件的发展而进行调优和配置。

为此，UEC考虑在链路层采用端到端基于信用的流控机制（CBFC）来管理链路间帧的无损传输。CBFC机制用来替换基于优先级流量控制（PFC）流控。接收者周期性发送缓存空间给对端，发送者基于报文优先级和缓存大小发送报文。缓存空间也可以用于自适应路由选路，同时配合链路遥测技术，准确感知整个网络的不同链路空闲和拥塞程度，及时调度流量，快速响应链路拥塞。

4）链路遥测技术。获得理想的拥塞往往需要及时感知网络链路状态和拥塞程度，因此我们需要研究端到端遥测技术。使用该技术可以准确获得网络的拥塞情况，及时将链路拥塞信号反馈回发送端，从而实现更快的拥塞控制。无论是发送方还是接收方安排传输，现代交换机都可以通过快速传递准确的拥塞信息给调度器，促进响应式的拥塞控制，提高拥塞控制算法的响应速度和准确性。链路遥测技术减少了拥塞，降低了丢包率，缩短了队列长度，降低了尾部延迟。

此外，网络系统可通过扩展LLDP协议，方便网络设备之间协商各自支持的链路层功能。这些功能包括超以太网技术中提及的新链路层功能，如LLR、CBFC、PFC等。

2.4 软件层技术

软件层工作组除了利用现有通信库设计开发各类应用通信接口和数据结构外，还研究在网计算相关工作，包括但不限于：1）基于C语言定义在网计算（INC）所使用的软硬件交互API接口；2）描述和定义硬件在网计算能力以及软硬件关于卸载能力的协商机制；3）设计和定义相关库函数、API接口实现主机与网络节点的数据交互，调用网络节点计算资源；4）OpenConfig扩展，用于配置网络设备的前端处理器（FEP）进行集合通信卸载，并对性能和错误进行监控；5）INC在网络设备上的适配，根据INC功能特性设计配置文件，并引导UEC传输协议的开发，以便INC技术可以轻松地应用到硬件实现中。

超以太网在链路层、传输层涉及的关键技术已在业界有了相关的研究，例如链路层的报文喷洒[11]、传输层的拥塞控制[8–10]，以及软件层涉及的在网计算[12]。超以太网技术与现有以太网技术不同的是，其主要面向AI计算中分布式资源的高效数据交互，即高带宽、低延时（低平均延时、低尾延时）传输需求。为此，超以太网技术可以借鉴现有的网络协议和相关技术研究，包括但不限于可编程数据平面、可编程网络、网络虚拟化、智能拥塞控制、网络链路遥测等，并在此基础上针对应用场景的特点，设计更加高效的传输协议和技术。

3 超以太网技术在中国的相关研究

针对高性能智算需求，中国的相关企业、高校、研究机构也积极布局下一代以太网技术，成立高通量以太网联盟、人工智能算力网络推进联盟等。

3.1 高通量以太网联盟

为应对AI数据中心网络面临的挑战，阿里云与中科院计算所联合成立高通量以太网联盟，旨在利用现有的以太网生态，优化传统以太网技术，研究和定义新型以太网协议和规范，设计新型智算网络，满足AI数据中心网络对高性能和低传输延时的需求。截至目前，高通量以太网联盟已经集结了大量中国学术界知名大学、产业界各类厂商和机构，打通理论研究、试验验证、产品部署全链条。

高通量以太网联盟在2024年计算机学会高性能计算学术年会上，对外发布了高通量以太网（ETH+）协议规范（1.0版本）、基于ETH+协议的相关开源网卡等硬件和系统。高通量以太网ETH+协议通过优化以太网帧格式，有效提升以太网帧的有效载荷比（74%），大幅提高AI数据中心大量短数据报文的传输效率。此外，ETH+以太网在链路层、物理层配套设计报文重传机制，有效提升数据传输的可靠性。与此同时，ETH+还可以支持在网计算功能，将原先在单节点上实现的部分计算卸载到网络节点中实现，可有效提升集合通信性30%以上的性能，从而解决传统网络单节点计算所存在的通信、计算瓶颈问题。

与超以太网联盟组织架构不同，高通量以太网联盟成立了协议标准和产业项目两个工作组。其中，协议组设计的高通量以太网协议和规范设计能够兼容现有以太网协议，并解决传统以太网协议可扩展性不足、负载不均、性能欠佳等问题。产业项目工作组负责针对差异化应用场景，将高通量以太网协议、规范应用部署其中，并负责项目实施落地工作。同时，联盟特设产业咨询会，负责跟进产业需求、拉动产业资源；设置执行小组制定技术路线图，协同推进各小组工作，从而促进中国各个芯片公司之间的合作与交流，推动技术创新和成果转化。

3.2 人工智能算力网络推进联盟

随着人工智能技术的迅猛发展，人工智能模型规模愈发庞大，原先一些小模型逐渐消失，取而代之的是“大模型+大数据+大算力”的紧密配合模式。从2018年的GPT到现在的GPT–4，大模型对算力的需求呈现指数增长态势，传统实验室、小型数据中心提供的算力已无法满足需求。

人工智能算力中心作为智能时代的新型公共基础设施，是人工智能产业发展的基础资源保障。为发挥其公共基础设施作用，必须要构建能够支撑人工智能产业持续发展的新型管理运营机制。为了促进中国战略性新兴产业的迅速发展和繁荣壮大，发挥各行业各地方在推进人工智能技术和产业发展的积极性，在鹏城实验室的倡议和推动下，“人工智能算力网络推进联盟”（简称“智算网络联盟”[13]）成立。

智算网络联盟目前已经有鹏城实验室、华为、百度、讯飞、燧原、天数智芯、北京智源研究院、武汉智算中心、珠海横琴智算中心等近20家单位参与。智算网络联盟将会在“平等自愿、优势互补、资源共享、合作共赢”的基础上，诚挚邀请致力于推动中国人工智能算力中心发展的企事业单位、科研院所、投资机构等加入。联盟成立后将重点在“智算中心及智算网络标准的研究及标准化”“推进成立人工智能算力网络管理中心”“组织开发并建设算力网络管理信息系统”“打造品牌活动，拓展影响”4个方面开展工作，致力于构建具有中国特色的新一代信息基础设施。

中国高通量以太网联盟、人工智能算力中心等联盟的成立，与超以太网联盟、开放计算项目联盟有着相似的目标，即针对AI计算对算力需求指数增长趋势，通过优化现有以太网技术、数据中心计算架构来实现通信能力和算力提升。

4 结束语

以太网凭借其高传输带宽、低成本、随即接入能力和成熟的协议生态，已然成为数据中心和高性能计算中心内部互连互通的关键技术。然而，传统以太网的优化技术主要考虑传输带宽的提升，在传输延时方面仍存在缺陷。超以太网联盟则针对现有以太网技术存在的缺陷展开研究，对其进行优化而非彻底颠覆。这种方式使得超以太网技术更容易被现有数据中心、智算中心接受和使用。

考虑到超以太网技术目前仍在发展初期，还未形成统一的协议规范，加上以太网生态的复杂性，因此对协议标准、技术、应用进行升级是一个巨大工程。超以太网技术离真正落地部署还有较长的距离。我们认为，超以太网技术未来要取得成功不仅需要依靠技术革新，还需要构建开源开放的生态，正如博通副总裁VEKAGA所说“不会有一家公司提供所有GPU，也不会有一家公司提供所有互连解决方案”。因此，超以太网快速发展的重要途径应该是建立一个生态系统，由多个供应商提供加速器。这个生态系统的生存依赖于构建一个开放的、基于标准的、高性能的和具有成本效益的互连架构。我们可以借鉴RISC-V开源指令集的思路，制定超以太网或者高通量以太网技术中基础且必须支持的协议规范，允许各大数据中心、厂商根据差异化的应用场景自定义扩展自己的协议规范，以吸引更多厂商和机构加入其中，从而进一步推进以太网技术的落地部署。此外，超以太网联盟还必须重视商业应用所注重的低成本、低复杂度、互连互通，才有可能使得超以太网技术进一步延伸至AI计算甚至高性能计算领域。

参考文献

[1] ZHAO W X, ZHOU K, LI J Y, et al. A survey of large language models [EB/OL]. [2024-10-10]. https://arxiv. org/abs/2303.18223 v15

[2] FLORIDI L, CHIRIATTI M. GPT-3: its nature, scope, limits, and consequences [J]. Minds and machines, 2020, 30(4): 681-694. DOI: 10.1007/s11023-020-09548-1

[3] ACHIAM J, ADLER S, AGARWAL S, et al. GPT-4 technical report [EB/OL]. [2024-10-04]. http://splab.sdu.edu.cn/GPT4.pdf

[4] Ultra Ethernet Consortium. The new era needs a new network [EB/OL]. [2024-10-01]. https://ultraethernet.org/

[5] Ultra Ethernet Consortium. Overview of and motivation for the forthcoming ultra ethernet consortium specification [EB/OL]. [2024-10-01]. https://ultraethernet. org/wp-content/uploads/sites/20/2023/10/23.07.12-UEC-1.0-Overview-FINAL-WITH-LOGO.pdf

[6] ZHANG H L, GUO X, YAN J Y, et al. SDN-based ECMP algorithm for data center networks [C]//Proceedings of IEEE Computers, Communications and IT Applications Conference. IEEE, 2014: 13-18. DOI: 10.1109/comcomap.2014.7017162

[7] ZHU Y B, ERAN H, FIRESTONE D, et al. Congestion control for large-scale RDMA deployments [C]//Proceedings of the 2015 ACM Conference on Special Interest Group on Data Communication. ACM, 2015: 523-536. DOI: 10.1145/2785956.2787484

[8] ALIZADEH M, GREENBERG A, MALTZ D A, et al. Data center TCP (DCTCP) [C]//Proceedings of the ACM SIGCOMM 2010 conference. ACM, 2010: 63-74. DOI: 10.1145/1851182.1851192

[9] KUMAR G, DUKKIPATI N, JANG K, et al. Swift [C]//Proceedings of the Annual Conference of the ACM Special Interest Group on Data Communication on the Applications, Technologies, Architectures, and Protocols for Computer Communication. ACM, 2020: 514-528. DOI: 10.1145/3387514.3406591

[10] MITTAL R, LAM V T, DUKKIPATI N, et al. TIMELY: RTT-based congestion control for the datacenter [J]. ACM SIGCOMM computer communication review, 2015, 45(4): 537-550. DOI: 10.1145/2829988.2787510

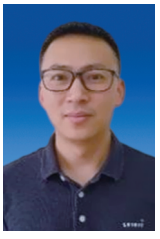
[11] ADDANKI V, GOYAL P, MARINOS I. Challenging the need for packet spraying in large-scale distributed training [EB/OL]. [2024-10-05]. https://arxiv.org/abs/2407.00550v1

[12] TOKUSASHI Y, DANG H T, PEDONE F, et al. The case for in-network computing on demand [C]//Proceedings of the Fourteenth EuroSys Conference 2019. ACM, 2019: 1-16. DOI: 10.1145/3302424.3303979

[13] 人工智能算力网络推进联盟 [EB/OL]. [2024-10-01]. https://c2net.openi.org.cn/

作 者 简 介

厉俊男，国防科技大学第六十三研究所助理研究员；研究方向为可编程网络处理器、低功耗嵌入式处理器；参与“863”计划、国家自然科学基金、武器装备预先研究等多项项目；发表论文10余篇，出版专著1部。



李韬，国防科技大学计算机学院网络空间安全系副研究员；研究方向为高性能网络芯片及系统；主持和参与“863”、重点研发、自然科学基金、武器装备预研等项目10余项，主持研制5款专用网络芯片；研究成果获4项科研成果奖；发表论文40余篇，出版专著2部，获授权专利20余项。

杨惠，国防科技大学计算机学院网络空间安全系副研究员；研究方向为高性能网络体系结构、网络处理器芯片；主持和承担芯片型谱、武器装备预先研究、重点研发、自然科学基金等国家及军队级项目10余项；发表论文30余篇，出版专著1部，获授权专利30余项。

基于生成式人工智能的算力网络自智优化研究综述

Self-Intelligent Optimization of Computing Power Networks Based on Generative Artificial Intelligence: A Review

崔佳怡/CUI Jiayi[1]，谢人超/XIE Renchao[1,2]，
唐琴琴/TANG Qinqin[1]
（1. 北京邮电大学网络与交换全国重点实验室，中国 北京 100876；
2. 紫金山实验室，中国 南京 211111）
(1. State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China;
2. Purple Mountain Laboratories, Nanjing 211111, China)

DOI：10.12142/ZTETJ.202406009
网络出版地址：http://kns.cnki.net/kcms/detail/34.1228.TN.20250106.1155.002.html
网络出版日期：2025-01-07
收稿日期：2024-10-25

摘要：生成式人工智能（GAI）技术可以在多样化业务处理过程中赋予算力网络（CPN）精准的意图分析能力，通过算网专家库的构建进而辅助算力网络实现高效的自适应智能决策，通过模型微调技术使资源配置决策适应突发网络变化，为用户提供精准且稳定的服务。基于上述目标，首先介绍生成式人工智能和算力网络概述，然后讨论了基于生成式人工智能的网络自智优化相关研究进展，创新性提出生成式算力网络的架构，对其核心流程和所需关键技术进行讨论，并对所提架构的优越性进行仿真验证和分析，最后对生成式算力网络应用场景进行分析，期望对该领域的后续研究提出可供借鉴的新思路。

关键词：生成式人工智能；算力网络；意图分析；模型微调

Abstract: Generative artificial intelligence (GAI) technologies can endow the computing power networks (CPNs) with precise intent analysis capabilities in diverse business processing scenarios. By constructing an expert database within the CPNs, it assists in achieving efficient adaptive intelligent decision-making. Through model fine-tuning techniques, the resource allocation decisions can adapt to sudden network changes, providing users with accurate and stable services. Based on these objectives, this paper firstly introduces an overview of GAI and CPNs, then discusses the research progress on network self intelligence optimization based on GAI. A novel architecture for generative computing power networks is proposed, along with discussions on its core processes and necessary key technologies. Furthermore, the superiority of the proposed architecture is validated and analyzed through simulation. Finally, an analysis of the application scenarios of generative computing power networks is provided, aiming to propose new perspectives for subsequent research in this field.

Keywords: generative artificial intelligence; computing power network; intentional analysis; model fine-tuning

引用格式：崔佳怡, 谢人超, 唐琴琴. 基于生成式人工智能的算力网络自智优化研究综述 [J]. 中兴通讯技术, 2024, 30(6): 54-62. DOI: 10.12142/ZTETJ.202406009
Citation：CUI J Y, XIE R C, TANG Q Q. Self-intelligent optimization of computing power networks based on generative artificial intelligence: a review [J]. ZTE technology journal, 2024, 30(6): 54-62. DOI: 10.12142/ZTETJ.202406009

随着人工智能的快速发展，越来越多的领域开始应用人工智能这一技术，如自然语言处理、计算机视觉等领域。在众多新兴人工智能应用中，生成式人工智能（GAI）作为其中的一个重要分支，在近年来取得了迅猛发展，它能够在几秒钟内生成高质量的内容，并根据用户的需求提供个性化的内容[1]。在新型人工智能技术的支持下，算力网络多种基础功能如任务分配、数据存储、计算处理等方面得到进一步优化，算力网络的应用场景也得到了不断拓宽。当前对于算力网络的研究正处于与新兴技术广泛融合的关键时期，生成式人工智能将促进算力网络的进一步发展，该技术可以自动化部署和管理算力网络中的异构资源，例如根据不同的任务特性和资源状态，动态地分配计算资源，优化计算路径，提高算力网络的运行效率和性能[2]。凭借对意图的精确感知和对海量数据的分析能力，生成式人工智

基金项目：国家自然科学基金项目（92367104）

能可应对多元化的算力场景和复杂化的业务需求，在匹配用户多维度需求方面为算力网络提供了更加智能高效的网络服务策略定制方案。

当前中国正在积极推动生成式人工智能和算力网络相关建设。2023年2月，中共中央、国务院发布《数字中国建设整体布局规划》，系统优化算力基础设施布局，促进东西部算力高效互补和协同联动，引导通用数据中心、超算中心、智能计算中心、边缘数据中心等合理梯次布局。2023年7月，中央网信办等七部门发布《生成式人工智能服务管理暂行办法》，首次明确了各方面的法定责任及法律依据，确立了人工智能产品的安全评估规定及管理办法。总之，数字经济已成为驱动中国经济发展不可或缺的力量，“网络赋能智能，智能使能网络”的创新发展已经逐渐成为推动国家数字化转型的重要力量。

本文通过对生成式人工智能和算力网络的系统调研，提出一种面向用户意图的生成式算力网络架构，并对其核心流程和关键技术进行探讨，随后进行仿真验证分析，最后对应用场景和未来发展方向进行分析展望，以响应生成式人工智能服务快速增长的算力需求，优化算力网络的整体机制，推动智能和网络的高效融合。

1 生成式人工智能与算力网络研究现状

在研究面向意图的生成式算力网络进行前需要明确相关概念，因此本节对生成式人工智能和算力网络相关概念和研究现状进行简要介绍。

1.1 生成式人工智能与算力网络概述

1）生成式人工智能

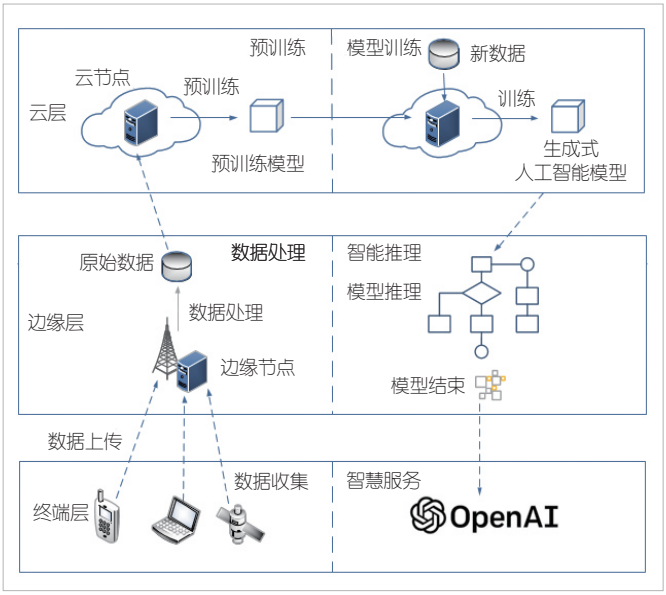
生成式人工智能是一种利用人工智能算法创造性地生成、操纵和修改有价值及多样化个性化数据的自动化方法[3]。生成式人工智能提供信息的过程不需要用户参与。在AI模型训练完成后，用户只需提供任务描述等输入，即可高效获取生成的内容。因为超高的生产内容效率，生成式人工智能逐渐成为新型网络的重要支撑工具。本节将介绍生成式人工智能最典型的服务架构和关键技术。

生成式人工智能架构以云–边–端三层网络架构为主。提供生成式人工智能服务的流程主要分为6个部分，包括数据收集、数据处理、模型预训练、模型训练、智能推理、智慧服务，如图1所示。该架构的核心在于根据任务的复杂性和数据规模，选择合适的模型架构来进行预训练，随后用特殊场景下的数据的分布规律对预训练模型进行再次更新，这个过程称为“微调”。生成式人工智能使用的预训练模型主要包括变分自编码器、生成对抗网络、扩散模型、Transformer等以及在这些基本模型上进行改进的架构。

生成式人工智能最初的应用面向人类与计算机的交互，针对人类需求自动化生成多媒体应用服务，如文章、音乐和图片等内容。随着研究范围的进一步扩大，多模态生成式人工智能技术赋予了基础大模型针对不同学科的强大知识和理解能力。当前全球针对生成式人工智能在网络当中应用的研究也更加深入，利用基础大模型能够生成各类面向系统的内容，辅助网络进行规划设计或资源管理等，对网络管理模块进行智能升级。例如：在生成式人工智能驱动网络架构的研究中，文献[4]介绍了生成式人工智能利用提示词工程将下游任务和生成式人工智能知识相关联，通过不同的提示策略优化生成质量，并使用离散提示和连续提示实现系统自动化。文献[5]主要考虑使用生成式人工智能作为网络组件或者增强网络功能，利用网内多模态数据，通过大语言模型（LLM）训练实现面向网络和任务的专业生成能力，例如网络协议、网络配置的设计和资源分配、链路管理策略的设计等。在感知通信数据处理方面，文献[6]提出生成式人工智能也不断增强数据集的多样性，利用有限的真实数据提高人类活动检测的精度，且可以解决相位模糊、信号到达方向估算等复杂问题。在利用信号频谱数据的训练中，生成式人工智能方法的性能明显优于深度强化学习方法。

2）算力网络

算力网络的核心思想是将分布的计算节点连接起来，动态实时感知计算资源和网络资源状态，进而统筹分配和调度计算任务，形成一张计算资源可感知、可分配、可调度的网



▲图1　生成式人工智能服务架构

络，满足新业务新应用对算力的要求。凭借其泛在算力按需分配的特点，算力网络已经成为驱动各行各业变革的重要解决方案。

当前全球针对算力网络的学术研究主要围绕算力资源建模、感知和调度三大类问题。对于算力资源建模，算力网络需要解决的最基本问题是如何衡量底层异构的算力资源质量和大小，并对其进行合理的表征和度量。近年来算网资源度量领域的研究者们对网络各种资源进行了全面度量分析[7-8]，兼顾算力资源的基础性能以及算力节点的工作状态，但随着智能的发展，需要引入更加自动化、更加智能的方法来提升度量的精确性。对于算力资源感知来说，基础设施层的算力资源庞大且分散，算力资源需要对计算任务进行按需匹配，感知机制的发展使得广泛的算力能够得到充分的调配协同，但目前的感知机制仍然无法解决跨域跨层级异构算力的全面感知[9-10]。对于算力网络协同调度来说，网络需要合理地对任务进行分配以及动态地检测和平衡运行中的节点，根据计算任务的要求，结合实时的计算负载和网络状态条件，动态地将计算任务调度到最匹配的边缘计算节点，实现对算力资源的协同利用和调度。尽管当前的算网调度机制已经非常完善，但在面对算网突发情况时，例如大规模节点环境的变化，仍难以做出即时决策[11]。

1.2 生成式人工智能与网络自智优化

生成式人工智能在网络中有显著优化能力，体现在自动学习、生成和改进网络相关的组件，以实现更加精准的网络服务响应。全球的相关研究主要集中在通过大模型训练实现面向网络和任务的专业能力。本文主要介绍以下几个方面：

1）网络模型应用

基于大模型的微调能力可以面向多种任务场景训练出针对特定场景问题的生成式解决方案。近年来，大语言模型如GPT-4和Llama-3等逐渐应用到文本解析、对话生成等多种自然语言处理任务中。在生成式人工智能对于网络自智优化的场景中，基于大语言模型的语义分析、内容生成、上下文学习等能力，通过少量数据微调训练能够捕捉专业化的语义关系和复杂的数据模式，实现面向网络和任务的专业生成能力。基于大语言模型设计的无线网络大语言模型能够解决正交频分复用（OFDM）系统的功率分配问题、无线网络的频谱感知问题、网络协议理解问题等，不仅可以突破大语言模型在无线通信中遇到的固有局限，还显著提升了其处理无线通信问题的能力[12]。除无线网络大语言模型之外，越来越多的定制化大模型逐渐支持6G客户端业务，通过数据、知识驱动的分布式协同部署和微调适配，在边缘侧充分发挥基站、边缘云的潜力，实现定制化大模型支撑多样个性化6G客户端业务。

2）网络自主设计

随着大规模网络的发展，第3代合作伙伴计划（3GPP）、美国电气电子工程师学会（IEEE）和国际电信联盟（ITU）等组织发布各种协议、标准和规范来确保设备之间的可靠高效通信。然而，协议的多样化和复杂性增加了网络对协议应用的困难性，并且制订的网络协议往往缺少自适应匹配网络环境的能力。生成式人工智能模型凭借突出的数据理解能力，可以快速获取与无线网络协议相关的信息。因此，利用生成式人工智能模型可以设计出更加智能的路由协议，根据网络流量的变化自动调整路由路径，以优化带宽使用和减少延迟，增强系统整体的链路性能[13]。除此之外，在大型网络环境中手动配置设备是非常耗时且容易出错的工作，借助生成式人工智能模型可以将这一过程变得自动化、智能化。系统可通过学习最佳实践和历史配置数据来生成合适的配置文件。

3）网络自智操作

生成式人工智能模型能够借助相关技术分辨和模拟复杂数据模式，更加智能地感知预测网络状态。该技术广泛应用于网络资源分配、链路管理策略生成等网络自智操作场景。因此，可以利用生成式人工智能模型学习网络历史流量模式，并预测未来流量，从而提前做出路由决策，自适应地分配带宽生成转发策略，甚至在实际数据缺乏时填补数据空白[14]。在某些应用场景下，比如视频流媒体或在线游戏，网络延迟和丢包率需要严格控制。对此，生成式人工智能模型可以用来优化服务质量（QoS）参数，确保优先级较高的流量获得更好的服务保障。除了这些资源分配策略生成的应用外，生成式人工智能还能够根据不断变化的网络条件和用户行为调整激励机制。例如，在混合现实（MR）场景中，生成式人工智能与契约理论等技术的结合能有效激励全双工设备对设备语义信息共享，避免重复的计算任务，以解决计算能力受限的问题。

2 生成式算力网络方案设计

为了应对算力网络多种应用场景下用户意图的差异化和个性化带来的新挑战，基于生成式人工智能技术和算力网络的研究现状，本节中我们提出面向用户意图的生成式算力网络，并给出生成式算力网络的定义。生成式算力网络在算网深度融合的基础上，更加关注用户意图的适配和算网环境的敏感变化，增强算网决策的智能生成能力和灵活适应能力，旨在精准满足用户意图的同时保障服务过程的稳定性，包括决策生成、模型更新、环境突变响应等功能。本节将介绍生

成式算力网络的设计动机，以及生成式算力网络的基础架构、核心流程和关键技术。

2.1 网络架构

生成式算力网络旨在根据生成式人工智能的经验、对算力网络中的当前基础设施资源状态和用户意图的感知，动态生成算力网络中网算存资源分配的最佳策略，实现“意图–策略”的高效匹配，同时实现网络内资源的灵活调度和协同，保障其全生命周期的安全性和可靠性，以提供高质量高可靠的生成式人工智能应用服务。生成式人工智能融合算力网络架构设计为3个层面，如图2所示，包括基础设施层、感知决策层、应用服务层，每一层实现功能具体如下：

1） 基础设施层

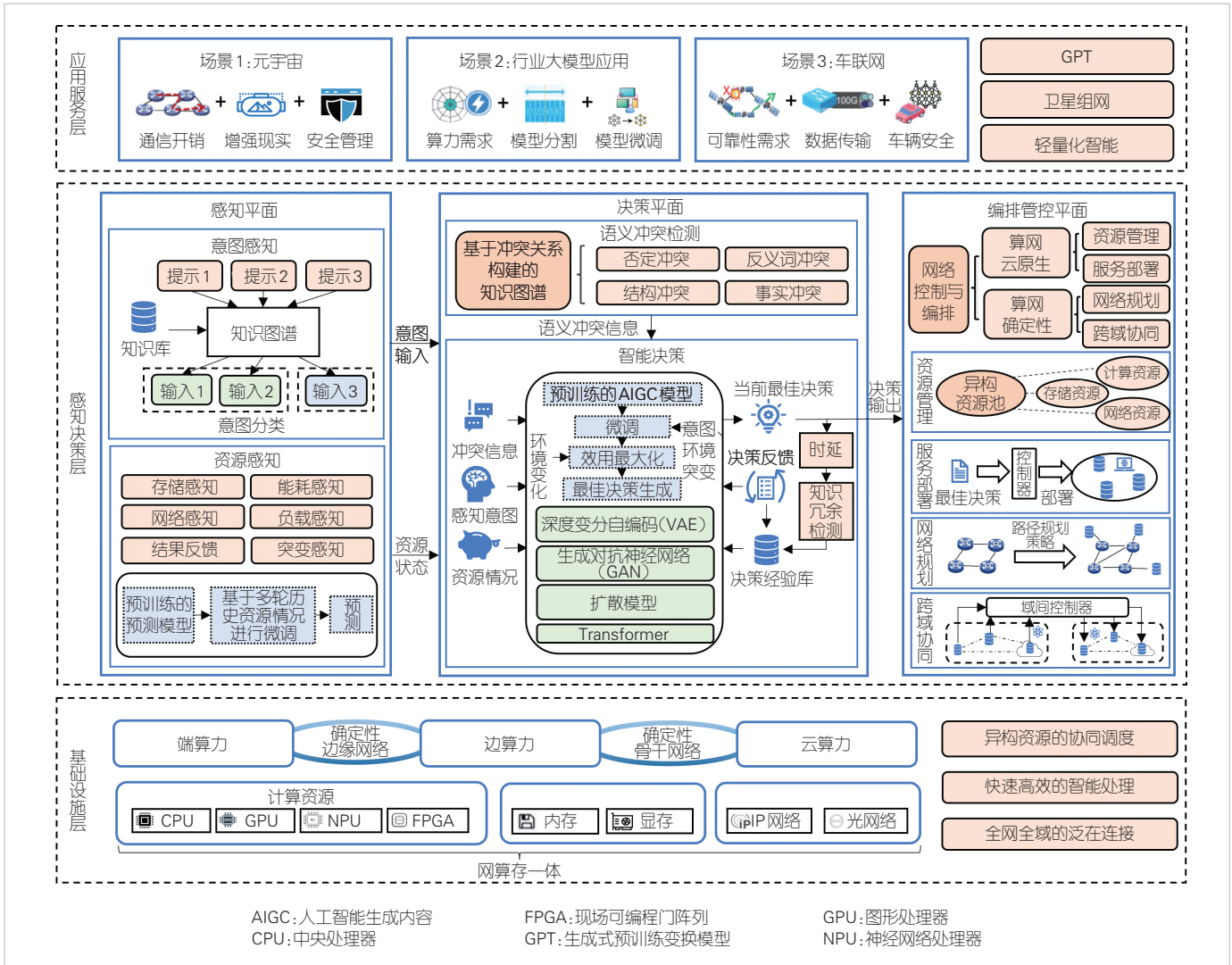
基础设施层为生成式算力网络提供了全网广泛的网算存资源。基础设施层中包含分布式算力节点和广泛的计算资源，算力节点包括端算力节点、边算力节点和云算力节点，这些算力节点部署基础算力、智能算力、超级算力等多种计算资源。确定性网络连接算力节点，实现全网全域的泛在连接。生成式算力网络将网络内的网算存资源整合统一，协同调度异构资源来高效地协同处理不同的任务，更好地响应用户需求，具备快速高效的智能处理能力。

2）感知决策层

感知决策层是生成式算力网络架构的核心。根据该层的功能，该层可分为3个平面：感知平面、决策平面、编排管控平面。

（1）感知平面

感知平面部分接收来自基础设施层和应用服务层的信息。感知平面从应用服务层获取用户的意图，用户意图往往



AIGC：人工智能生成内容　　FPGA：现场可编程门阵列　　GPU：图形处理器
CPU：中央处理器　　GPT：生成式预训练变换模型　　NPU：神经网络处理器

▲图2　面向用户意图的生成式算力网络架构

包含于语义信息中，感知平面主要应用基于知识图谱的意图转译方法。除此之外，感知平面实时接收基础设施层上传的资源状态，包括“网、算、存”资源的使用情况和负载均衡情况，以便后续根据用户意图和资源情况进行协同考虑，在完成最优决策部署的同时最大化资源的利用率。为了灵活应对环境资源产生的多种敏感变化，提高生成决策性能的稳定性，感知平面对资源的感知不能只停留在当前时间状态，而是需要根据历史资源状态的存储记录对未来某一时刻资源状态进行预测，以适应决策生成过程中时延带来的环境改变。同时，感知平面也感知任务的执行状态，检查决策是否执行以及其性能优劣。此外，感知平面还具备监测环境突变（如基础设施节点大范围的减少或增加）的能力，及时发现对用户服务不利的情况以做出灵活应对，保持决策的可用性，提供服务的持久稳定性。

（2）决策平面

决策平面是整个生成式算力网络工作的关键，其核心能力是使用生成式人工智能相关技术，基于感知平面输入的用户意图、接收到的当前和未来预测资源生成最优策略，为用户目标定制个性化资源调度策略。决策平面引入“以意图为中心”的算网专家库，它相当于一个历史记录存储库，记录生成式算力网络中所有的用户意图部分信息和与之匹配的最终决策模型的参数。利用用户意图之间的相似性，算网专家库为新用户意图提供有效的预训练生成式人工智能模型选择空间，减少重新训练模型的时间。算网专家库辅助选择的预训练生成式人工智能模型，不能够完全满足用户个性化的意图需求，因此需要根据用户意图对预训练模型进行适应性微调，使模型的性能和损失逐渐符合用户的期望，同时通过不断与资源环境反馈状态进行实时交互，从而不断更新模型参数，更好地适应环境。除此之外，预训练生成式人工智能模型的微调不止发生在用户意图输入的时刻，还发生在环境大规模突变的时刻，以适应环境输入维度的变化，改变模型使其能够生成更加合理的决策，保障用户服务执行过程中不受环境突变的影响。

（3）编排管控平面

编排控制平面是决策的执行者，它根据最优决策对网算存资源进行一体化管理，部署最佳决策内容到基础设施层的节点执行，调度分配、管理算力资源和规划网络，同时控制跨域协作，从而可以实现多域算力共享和域间服务协作。

3）应用服务层

生成式算力网络的应用服务层不仅能够满足更多智能应用的需求，如元宇宙、GPT、车联网等，还能提供生成式人工智能服务，如文本、图像、视频等内容，并且支持内容根据用户意图自适应调整等功能，以不断满足用户新需求。

2.2 核心流程

针对生成式算力网络的核心部分，即决策平面，分析决策产生的主要工作流程，包括决策生成、模型更新和环境突变响应。

1）决策生成

决策生成流程根据当前资源状态生成最优策略。基础设施层进行语义分割，度量每个节点的资源状态，包括存储、能耗、网络、负载和历史策略执行性能，然后将输入传输到感知决策层。在感知平面上，基于知识图谱将语义分割生成的文本划分为不同的意图组，同时感知资源状态和决策反馈。接下来，决策平面根据知识图中不同单词之间的关系，检测意图组之间的语义冲突。同时，利用预测模型，根据之前的输入，预测未来短时间内资源的状态。将用户意图在经验库中进行搜索和近似匹配，调取预训练生成式人工智能模型。使用用户意图对预训练生成式人工智能模型进行小样本微调，然后将资源预测结果、环境冲突信息、用户意图、当前资源状态等信息反馈给核心预训练生成式人工智能模型生成最优决策，同时进行模型更新和环境突变响应以实现决策效用最大化。收到决策平面响应后，系统将决策发送到编排控制平面，控制器根据该决策管理部署节点，执行任务调度。最后，基础设施层中的节点接收部署的决策并执行，完成用户服务后反馈结果信息。

2）模型更新

模型更新过程包括两部分，这两部分分别根据来自应用服务层的用户意图和来自基础设施层的资源反馈进行模型参数的更新。用户意图对从算网专家库中选取的预训练生成式人工智能模型进行微调，使其更满足当前用户的需求，以获得用户意图的最佳匹配决策。这比直接使用预训练生成式人工智能模型更准确。同时，模型在决策生成与执行完成的过程中，不断与基础设施层反馈的资源环境以及执行过程中反馈的结果进行交互，并不断更新模型参数，使模型适应环境敏感变化，保障生成内容的有效性。执行完决策之后，算网专家库确认最终决策，将最终模型添加到专家库中，同时将时间较远的模型参数删除，以减少专家库的数据冗余。

3）环境突变响应

环境突变响应是一种发生于环境急剧变化的决策模型适应过程。发生环境突变主要有以下几种情况：（1）基础设施层中负载不均衡，造成大量节点过载无法使用，空闲节点没能得到充分利用；（2）网络规模的扩大和缩小，例如算力供应商的加入和退出造成大量节点的增删情况，资源环境维度

发生变化。针对第一种情况在基础设施层设置负载阈值，以判断整体网络环境内所有决策的综合性能。如果过载节点数大于阈值，感知决策层将接收到相关信息，将目前正在执行的决策终止，根据当前信息立即制定新的策略。针对第二种情况，需要对当前执行模型和算网专家库中的模型全部进行微调，使其迅速适应大范围的网络变化。

2.3 关键技术

本小节将从生成式算力网络架构和流程中的几种关键支撑技术展开探讨，包括意图感知和转译技术、预测和决策生成模型、微调技术等。

1）意图感知和转译技术

在基础设施层，使用包括解码器网络的语义分割将用户的需求划分为文本提示。然后，意图感知部分利用知识图谱对文本提示进行分类。知识图谱用于表示实体之间的关系和属性，以及其语义信息。首先基于开放关系抽取[15]等方法可以通过分析用户输入的语句，从非结构化文本中提取开放领域的关键实体、特征和关系信息，并将这些信息映射到知识图谱中相应的节点和边。目标本身固有的属性信息是用于目标意图分析的参数，如网算存资源参数和网算存节点等。以各种特征为基本节点的节点信息主要包括距离、能耗、资源利用率等特征参数，而进行意图分析的节点关系主要包括目标实体意图间的包含关系、关系值域、关系约束等。借助一系列的知识抽象、知识推理、构建上述知识图谱，并通过反馈实时更新数据库，系统可实现更加精确的目标意图知识图谱。利用知识图谱即可进行意图的分析和转译，获得意图的各种属性、特征和关系解析结果。

2）预测和策略生成模型

在生成决策前，系统可通过构建神经网络对未来资源状态进行预测，输入资源的当前状态并估计资源的后续状态，以减少时间延迟对策略性能的影响。决策生成模型主要基于生成式人工智能模型。生成式人工智能模型包括VAE、GAN、扩散模型、Transformer等经典内容生成模型，提供动态自适应调度方案[16]，提升系统的智能水平。除此之外，也可采用强化学习算法对流程进行优化，例如Actor-Critic结构和经验缓冲区。

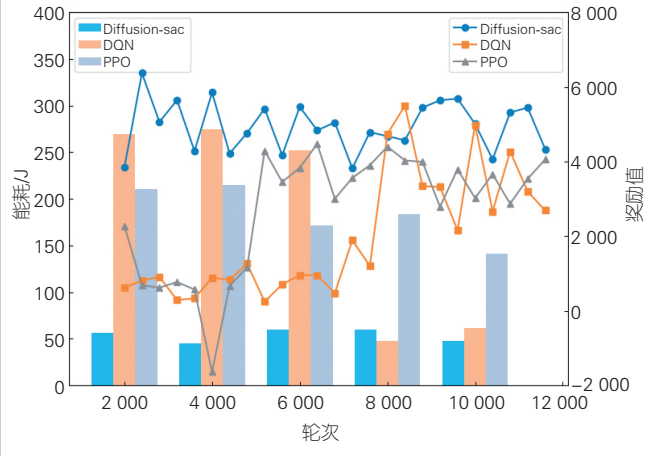
3）微调技术

参数高效微调是指微调少量或额外的模型参数，固定大部分预训练模型参数，从而大大降低计算和存储成本，同时也能实现与全量参数微调相当的性能。参数高效微调方法甚至在某些情况下比全量微调效果更好，可以更好地泛化到域外场景。高效微调技术可以粗略分为以下三大类：增加额外参数（例如AttentionFusion）、选取一部分参数更新（例如BitFit）、引入重参数化（例如LoRa）。其中，增加额外参数这类方法又主要分为类适配器（Adapter-like）方法和软提示（Soft prompts）两个小类[17]。如果环境中的突变对决策的性能产生较大影响，就需要使用微调技术来调整模型。此外，为了满足用户输入意图，对从算网专家库中复制的预训练模型也要进行微调，基于新的数据集更新预训练模型部分参数以适应新的任务。

2.4 仿真验证

本文中，我们通过仿真实验对生成式算力网络架构以及扩散模型集成的强化学习算法进行简单验证，并对3种算法的奖励值进行仿真：集成扩散模型的强化学习算法（Diffusion-sac）、深度Q网络（DQN）和基于策略梯度的强化学习算法（PPO）。对生成式算力网络的智能生成策略机制进行仿真验证主要分为两部分：1）选择每轮训练的奖励值和网络能耗作为系统性能指标，验证生成式算力网络和普通算力网络的策略生成性能；2）对不同用户算力需求下系统的表现性能和稳定性进行仿真时延，以证明其在不同场景下的适应能力。我们使用扩散模型作为Actor网络核心算法，对于实时变化的网络环境和用户需求做出算力的分配策略，关注3种学习方法的奖励值曲线和网络整体的能耗指标。

在策略生成性能方面，如图3所示，Diffusion-sac算法学习奖励值的曲线平稳且高于其他算法，在任务处理能耗方面也使得网络始终具备最低能耗。这些结果证明了Diffusion-sac算法在高性能、低能耗和快速收敛方面的优越特性。因此，Diffusion-sac算法是解决算力网络场景策略生

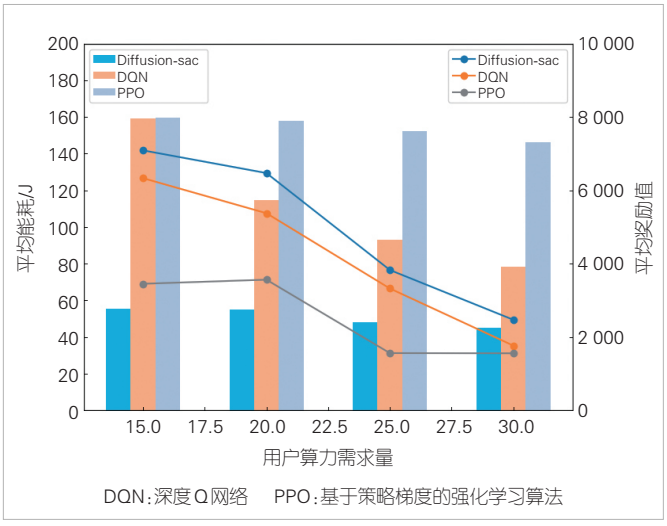


DQN：深度Q网络　PPO：基于策略梯度的强化学习算法

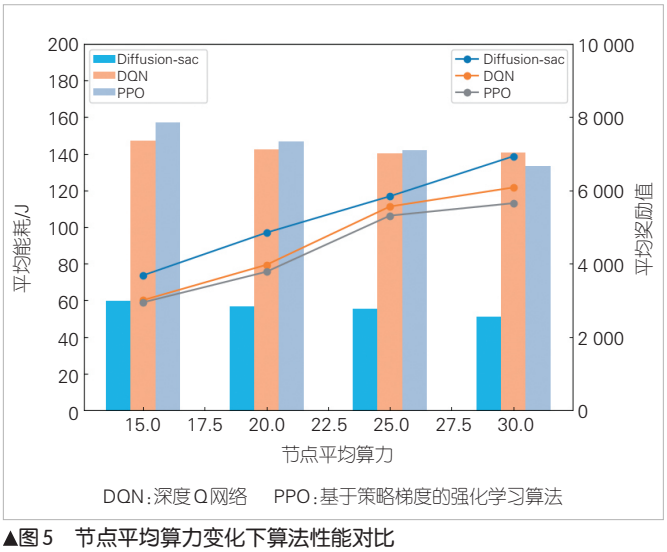
▲图3　3种算法奖励值和能耗对比

成问题的首选。

在系统多场景下的稳定性方面，图4和图5显示了在不同的用户算力需求以及不同节点算力容量情况下Diffusion-sac算法的稳定性。可以看出，随着用户算力需求量的增加，3种算法的平均奖励值有所下降，这是因为节点资源承受的资源供给压力较大，但是Diffusion-sac算法始终保持着最高的平均奖励值，并且能耗值也远小于其他算法。由节点平均算力的增加可以看出，3种算法的奖励值都有所提升。与此同时，系统的整体平均能耗有所下降，这表明系统具备负载均衡的能力，不会出现能耗过高情况，而且Diffusion-sac算法仍然保持着最优效果。实验证明，Diffusion-sac算法不仅在单一场景下具备优秀的训练能力，并且在变化的负载条件下也能够适应场景的变化进行最优决策，具备强大的泛化能力和稳定性能。



▲图4　用户算力需求量变化下算法性能对比



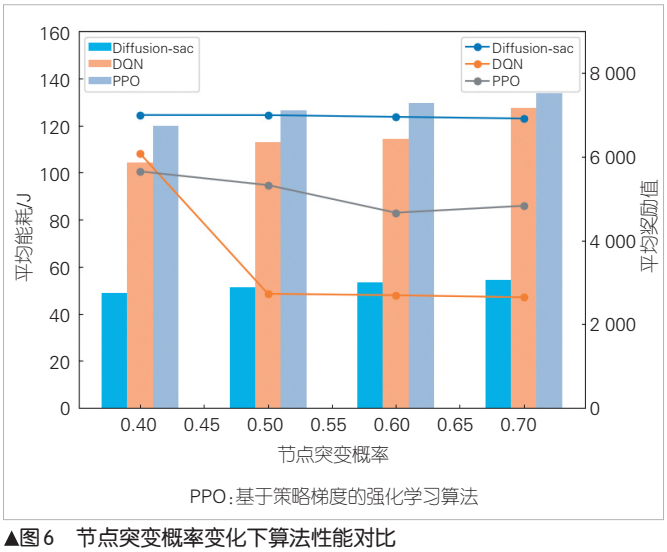
▲图5　节点平均算力变化下算法性能对比

除此之外，我们还进行了针对模型泛化能力的仿真验证。图6展示了改变环境节点突变概率时3种算法的表现情况。我们设置节点突变概率变量，用于控制算网环境中每个节点的崩溃概率。若节点发生崩溃，则任务调度过程中需要考虑其他节点情况。其中，3种算法的奖励值整体趋势都随着节点突变概率的提升而下降。这是因为可用节点数量越少，任务处理和调度决策困难就越会导致节点负载不均匀，奖励值就越下降。由图6可知，Diffusion-sac算法的奖励值变化幅度较小，且始终高于其他两种传统算法，这显示了Diffusion-sac算法的稳定性能。3种算法产生的能耗整体趋势都随着节点突变概率的提升而有所提升，这是因为节点负载不均衡导致部分节点使用过载，能耗极具增大。其中，能耗最小的算法为Diffusion-sac，在每种情况下都表现为最小，这说明该算法能够尽可能在环境突变的情况下保持负载均衡的决策。

3 生成式算力网络应用场景

3.1 生成式算力网络服务于垂直行业

大模型的应用对于各专业场景都是新兴的发展趋势，但训练大模型对大多数企业来说不现实，因为这些企业往往只专注于一类业务，更希望利用大模型在行业场景的个性化智能能力。企业的需求是支付尽可能低廉的成本，而不是付出巨大算力直接进行大模型训练。总结来说，当前垂直行业应用生成式人工智能技术存在以下问题：1）由于大模型训练需要大量的算力资源，目前只有云端能够支撑大模型训练方案，边缘算力仅依靠协同计算无法支撑大模型训练，绝大多



▲图6　节点突变概率变化下算法性能对比

数用户和企业没有能力训练自己的大模型；2）垂直行业设备大量分布在边缘端，因此生成式人工智能的服务交互需要向云端请求并获得结果反馈，通信开销巨大；3）目前现有的大模型大多数是通用模型，不符合垂直行业的专业需求。

生成式算力网络为垂直行业训练属于自己的大模型以提供算力条件支撑。网络基础设施层结合多种分布式智能技术，使得预训练大模型在边缘为各专业领域提供个性化服务成为可能，大大减少了实现个性化智能能力的复杂性。文献[18]中基于给定主题的几张图像，将这几张图片的特征嵌入到模型的输出域中，并对预训练的基于扩散的文本到图像框架进行微调。这样的解决方案不仅能够满足企业训练模型需求，也保证了智能资源的充分使用，解决高昂的成本问题。

3.2 生成式算力网络实现服务个性化定制

传统算力网络根据业务需求，通过移动边缘计算等技术将计算资源、存储资源以及网络资源进行集中和灵活调度，以实现按需的算力分配和灵活调度。由于不同的业务需求需要不同的计算、存储和网络资源，而且同一业务的不同阶段也可能需要不同的资源分配，因此，算力网络需要提供个性化的服务，以满足不同业务需求和用户需求，同时对资源进行高效利用，避免资源的浪费。但随着智能业务的海量涌现和发展，传统的静态资源管控调度方案面临灵活性差、适配度低等问题。

在生成式算力网络中，生成式人工智能算法会根据用户的行为和偏好，生成个性化的服务内容和推荐。例如，网络可以根据用户的浏览历史、购买记录、搜索关键词等信息，生成个性化的推荐列表，帮助用户更快速地找到自己感兴趣的内容。同时，算力网络还可以根据用户的反馈和评价，不断优化和改进个性化服务。例如，如果用户对某个推荐的服务或商品不满意，融合网络可以根据用户的反馈信息，重新调整算法模型，提高个性化推荐的准确性和质量[19]。

3.3 生成式算力网络驱动虚拟世界扩展

元宇宙作为未来虚拟世界的起点已经获得了极大的关注，然而数据的映射是在虚拟世界和现实世界之间建立共生互联网的先决条件。生成式人工智能技术通过利用人工智能的力量来自动化信息创建过程，为快速创建数字内容提供了技术支持。但随着虚拟世界的逐渐扩展和演进，虚拟世界用户数据量的增多带来大模型广泛访问的问题。生成式算力网络的出现能够为大模型提供缓存空间，解决访问频繁问题。

通过生成式算力网络可以实现跨平台一体化的元宇宙体验。无论是对于虚拟现实设备、增强现实设备，还是针对智能手机等移动终端，算力网络承载的生成式人工智能都可以提供一致且无缝的用户体验，使用户可在不同设备上连续获得元宇宙服务。算力网络不仅能为用户提供娱乐环境，还能帮助用户实现元宇宙的任务管理、资源调配、时间规划等，进而提高用户的工作效率[20]。

4 生成式算力网络未来发展

4.1 激励机制

在生成式算力网络中，生成式人工智能服务的整个生命周期需要对参与者进行适当的激励。由于网络中参与生成式人工智能服务的节点提供异构资源，在提供服务的过程中，数据收集、预训练、调优和推理都需要大量异构资源参与。这就需要对服务的各方参与者按照其贡献设计合理的激励机制，例如引入区块链技术，实现网络对生成式人工智能服务全生命周期的去中心化管理。用户可以根据服务提供商的交易历史来评估其声誉，从而促进服务的优化和改进。

4.2 服务定制

在生成式算力网络中，生成式人工智能模型的预训练、微调和推理通常会消耗大量的计算和网络资源。因此，我们可以更加关注生成式算力网络的绿色运营，以最小的能耗和碳排放提供生成式人工智能服务。此外，还可以提出智能资源管理和调度技术来平衡服务质量和资源消耗。

4.3 模型压缩

随着生成式人工智能大模型变得越来越复杂庞大，在提供生成式人工智能服务时，模型压缩技术对于减少服务延迟和资源消耗变得越来越重要。目前已经研究出的模型压缩技术包括：修剪、量化和知识蒸馏技术。其中，修剪技术的目的是去除不重要的权重，而量化则降低了权重的精度，知识蒸馏目的是训练一个较小的模型来模仿较大模型的行为。未来对于模型技术的研究可以基于上述技术进行完善，以平衡模型的大小和精度。

4.4 服务安全隐私

为了提供保护隐私的生成式人工智能服务，在模型训练和推理中都需要考虑隐私计算技术。差分隐私、安全多方计算和同态加密等技术可用于保护敏感数据并防止未经授权的访问。其中，差分隐私涉及在数据中添加噪声以保护个人隐私，安全多方计算允许多方在不向彼此透露其输入的情况下计算一个函数，同态加密允许在不解密的情况下对加密数据

执行计算。

5 结束语

由于生成式人工智能对用户需求高效快速响应，在任务处理方面带来显著的性能提升，同时算力网络能够为生成式人工智能提供非常充足的计算条件，因此在算力网络中引入生成式人工智能的新型网络架构具备充分的条件。本文对算力网络的概念进行详细介绍，并系统梳理了生成式人工智能的发展现状，重点对生成式人工智能融合算力网络的系统架构、通信流程及关键技术进行概述，最后通过仿真实验验证了生成式算力网络架构的合理性和有效性。在未来的工作中，我们可以对生成式算力网络架构每一层进行详细的研究，以期提供更加完善高效的智能算力服务。

参考文献

[1] LEE L H, LIN Z J, HU R, et al. When creators meet the metaverse: a survey on computational arts [EB/OL]. [2024-10-25]. http://arxiv.org/abs/2111.13486

[2] 彭开来, 王旭, 唐琴琴. 算力网络资源协同调度探索与应用 [J]. 中兴通讯技术, 2023, 29(4): 26-31. DOI: 10.12142/ZTETJ.202304006

[3] HARSHVARDHAN G M, GOURISARIA M K, PANDEY M, et al. A comprehensive survey and analysis of generative models in machine learning [J]. Computer science review, 2020, 38: 100285. DOI: 10.1016/j.cosrev.2020.100285

[4] LIU Y Q, DU H Y, NIYATO D, et al. Optimizing mobile-edge AI-generated everything (AIGX) services by prompt engineering: fundamental, framework, and case study [J]. IEEE network, 2024, 38(5): 220-228. DOI: 10.1109/MNET.2023.3335255

[5] BARIAH L, ZHAO Q Y, ZOU H, et al. Large generative AI models for telecom: the next big thing? [J]. IEEE communications magazine, 2024, 62(11): 84-90. DOI: 10.1109/MCOM.001.2300364

[6] ZHENG J K, ZHANG J Y, DU H Y, et al. Flexible-position MIMO for wireless communications: fundamentals, challenges, and future directions [J]. IEEE wireless communications, 2024, 31(5): 18-26. DOI: 10.1109/MWC.011.2300428

[7] 杜宗鹏, 李志强, 陆璐. 算力网络四面三级算力度量技术体系 [J]. 中兴通讯技术, 2023, 29(4): 8-13. DOI: 10.12142/ZTETJ.202304003

[8] 李重严, 毕成, 张晟. 面向信息能源融合的低碳算力网络架构研究 [J]. 电信工程技术与标准化, 2022, 35(11): 1-6. DOI: 10.3969/j.issn.1008-5599.2022.11.001

[9] 闫实, 彭木根, 王文博. 通信-感知-计算融合: 6G愿景与关键技术 [J]. 北京邮电大学学报, 2021, 44(4): 1-11. DOI: 10.13190/j.jbupt.2021-081

[10] 许胜, 许方敏, 赵成林. 基于数字孪生的算力网络自优化技术研究 [J]. 中兴通讯技术, 2023, 29(3): 46-50. DOI: 10.12142/ZTETJ.202303009

[11] 衷璐洁, 王目. 区块链赋能的算力网络协同资源调度方法 [J]. 计算机研究与发展, 2023, 60(4): 750-762

[12] SHAO J W, TONG J W, WU Q, et al. WirelessLLM: empowering large language models towards wireless intelligence [J]. Journal of communications and information networks, 2024, 9(2): 99-112. DOI: 10.23919/JCIN.2024.10582827

[13] 任天骐, 李荣鹏, 张宏纲. 通信网络与大模型的融合与协同 [J]. 中兴通讯技术, 2024, 30(2): 29-36. DOI: 10.12142/ZTETJ.202402005

[14] LIU Y Q, DU H Y, NIYATO D, et al. Deep generative model and its applications in efficient wireless network management: a tutorial and case study [J]. IEEE wireless communications, 2024, 31(4): 199-207. DOI: 10.1109/MWC.009.2300165

[15] DAI H, DU D, LI X, et al. Improving fine-grained entity typing with entity linking [J/OL]. [2019-09-26]. https://arxiv. org/abs/1909.12079

[16] SU N. Research on multiparty participation collaborative supervision strategy of AIGC [C]//Proceedings of IEEE 13th International Conference on Electronics Information and Emergency Communication (ICEIEC). IEEE, 2023: 268-272. DOI: 10.1109/ICEIEC58029.2023.10200392

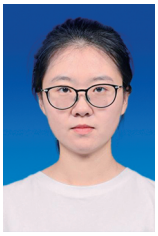
[17] 丁鑫, 邹荣金, 潘志庚. 基于高效参数微调的生成式大模型领域适配技术 [J]. 人工智能, 2023(4): 1-9

[18] RUIZ N, LI Y Z, JAMPANI V, et al. DreamBooth: fine tuning text-to-image diffusion models for subject-driven generation [C]// Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2023: 22500-22510. DOI: 10.1109/CVPR52729.2023.02155

[19] DU H Y, LI Z H, NIYATO D, et al. Enabling AI-generated content services in wireless edge networks [J]. IEEE wireless communications, 2024, 31(3): 226-234. DOI: 10.1109/MWC.004.2300015

[20] DU B X, DU H Y, LIU H F, et al. YOLO-based semantic communication with generative AI-aided resource allocation for digital twins construction [J]. IEEE Internet of things journal, 2024, 11(5): 7664-7678. DOI: 10.1109/JIOT.2023.3317629

作 者 简 介



崔佳怡，北京邮电大学在读博士研究生；主要研究领域为算力网络、工业互联网等。

谢人超，北京邮电大学教授；主要研究领域为未来网络体系架构、算力网络、云网融合、工业互联网、信息中心网络等；作为项目负责人主持或参与国家重点研发计划、国家自然科学基金、北京市自然科学基金、工信部重大专项、华为企业合作基金等项目20余项；发表论文70余篇。

唐琴琴，北京邮电大学博士后；主要从事边缘计算、算力网络、卫星互联网、网络人工智能相关研究工作；参与多个国家重点研发计划、国家自然科学基金等项目；发表论文20余篇，申请国家发明专利10余项。

HPN：阿里云大模型训练网络架构

HPN: Alibaba Cloud's Data Center Network Architecture for Large Language Model Training

钱坤/QIAN Kun，翟恩南/ZHAI Ennan，操佳敏/CAO Jiamin
（杭州阿里云飞天信息技术有限公司，中国 杭州 310030）
(Hangzhou AliCloud Apsara Information Technology, Hangzhou 310030, China)

DOI：10.12142/ZTETJ.202406010
网络出版地址：http://kns.cnki.net/kcms/detail/34.1228.TN.20250109.0925.004.html
网络出版日期：2025-01-09
收稿日期：2024-10-16

摘要：介绍了阿里云用于大型语言模型（LLM）训练的数据中心网络架构高性能网络（HPN）。HPN通过双上联、多轨、双平面的网络架构设计，避免了单链路故障带来的严重连通性影响，并且避免了哈希极化的产生。实验表明，HPN将LLM训练的端到端性能提升超过14.9%。HPN已在阿里的生产环境中部署了超过1年。

关键词：大模型训练；网络架构；数据中心网络

Abstract: The Alibaba cloud's data center network architecture for high-performance network (HPN) used in the training of large language models (LLMs) is introduced. HPN is designed with a dual-top of rank (ToR), rail-optimized, and dual-plane architecture, which avoids severe connectivity impacts caused by single-link failures and prevents hash polarization. Experiments have shown that HPN improves the end-to-end performance of LLM training by over 14.9%. HPN has been deployed in Alibaba's production enviroment for over a year.

Keywords: large-scale model training; network architecture; data center network

引用格式：钱坤, 翟恩南, 操佳敏. HPN：阿里云大模型训练网络架构 [J]. 中兴通讯技术, 2024, 30(6): 63-67. DOI: 10.12142/ZTETJ.202406010
Citation：QIAN K, ZHAI E N, CAO J M. HPN: Alibaba cloud's data center network for large language model training [J]. ZTE technology journal, 2024, 30(6): 63-67. DOI: 10.12142/ZTETJ.202406010

大语言模型（LLM）包含超过100亿个参数，并且由多个模型层构成。这些模型的高效训练需要数千个图形处理器（GPU）协同工作。主流的训练框架（例如Megatron-LM[1]和Deepspeed[2]）通常通过多种并行策略的混合来实现大规模训练。

1）数据并行（DP）。训练数据集均匀分布在所有GPU之间，每个GPU都拥有整个模型的一个副本。在每次迭代中，所有GPU都使用AllReduce来同步计算出的梯度。

2）流水线并行（PP）。模型被划分为多个阶段，每个阶段包含一系列连续的模型层，并由不同的GPU提供服务。流水线中的每个GPU都接收来自前一阶段的输入，并将输出发送到流水线中的下一阶段。

3）张量并行（TP）。在PP中，整个模型或每个层都可以进一步水平分割。因此，每个层都分布在一组GPU之间。在每次迭代中，同一TP组中的GPU使用AllReduce/AllGather来同步计算输出和相应的梯度。

考虑到大模型训练过程中的各种并行策略，训练过程中观察到的流量模式与弹性云计算或传统深度神经网络（DNN）训练中的流量模式非常不同，这样的流量模式的差异给智算集群网络带来了新的挑战。

1 AI大模型训练的网络挑战

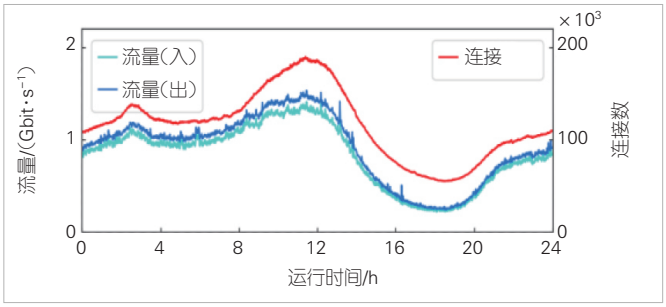
1.1 挑战1：AI训练流量特性导致负载均衡困难

传统的数据中心网络架构（例如fat tree[3]）主要用于一般的弹性云计算。我们观察到，LLM训练的流量模式与一般云计算有所不同。

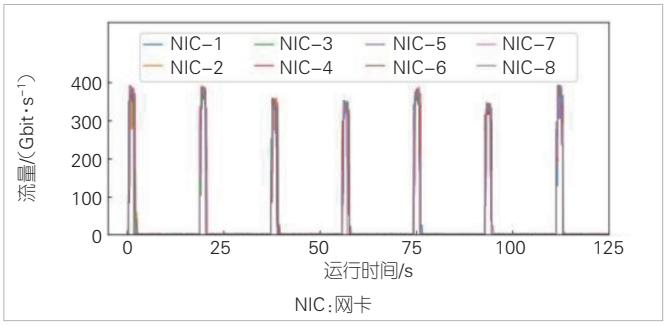
1）网络利用率周期性突发。在实际生产中，传统云计算会产生数百万个流量，流量利用率通常保持在20%以下。整体流量模式相对连续稳定，每小时缓慢变化（如图1所示）。

相反，LLM训练产生的连接极少，但流量存在周期性突发（如图2所示）。图2显示了我们实际生产中LLM训练中一块网卡（2×200 Gbit/s）的吞吐量。网卡周期性地传输大量数据，瞬间就会达到端口最大容量（400 Gbit/s），并会持续几秒到几十秒。在每个迭代中，需要在不同的并行组（每个具有许多GPU）之间进行数据同步。

网络利用率的突发意味着LLM训练需要极高的网络带



▲图1 传统云计算流量模式



▲图2 模型训练期间的网卡出方向流量

宽。因此，我们需要确保LLM训练的网络能够为突发情况提供足够带宽，以避免丢包。此外，流量的同步性表明，LLM训练对长尾延迟特别敏感。任何长尾流量都将阻碍整个集合通信操作的完成，这会使所有参与这次集合通信的所有节点陷入等待。

2）流量数量较少。如图1所示，一般的云计算实例通常会产生数十万个连接；相反，LLM训练中的每个节点产生的连接非常少，一个GPU仅使用几十到几百个连接。结合之前训练过程中提到的突发高网络利用率，每个流需要发送的实际数据量是相当大的。

3）负载不均。传统数据中心网络采用等价多路径路由（ECMP）作为负载均衡方案。ECMP假设哈希算法能够有效地将流量均匀地分布在网络中所有等效路径上，当网络中有大量流数目时，此假设是成立的。然而，在LLM训练中，这种假设不再成立，因为LLM训练仅涉及少量大流。在我们使用传统数据中心进行LLM训练的实践中，已经遇到了很多由此引起的性能问题。

更严重的是，由于传统数据中心网络都采用了3层网络架构，大流的转发需要经过3次哈希计算（即柜顶交换机、聚合和核心层）。由于每次哈希的输入（即流的五元组）保持不变，这种“级联”哈希的影响会导致更严重的负载不平衡（即哈希极化[4]）。我们在生产中也确实观察到了很多由于哈希极化而引起的负载不均的现象。这样的问题在跨Pod通信场景中尤为常见。

1.2 挑战2：AI大模型训练对网络故障更敏感

1）LLM训练对故障更为敏感。在LLM训练中，多个GPU合作完成每个迭代，并且我们需要许多次迭代（持续几十天）来完成整个训练过程。因此，任何一个GPU或主机的故障都可能直接导致整个LLM训练过程崩溃。

2）预防单点故障很重要。传统网络架构中尽管Tier2和Tier3层具有丰富的冗余链路，但每个网卡（NIC）只通过一条链路连接到柜顶交换机（ToR），存在单点故障风险。当接入链路（即连接NIC和ToR的链接）中断时，对应的主机会出现连接断开的情况。更糟糕的是，ToR的故障可以使数十甚至数百台主机不可用，这会导致严重的服务质量下降。LLM训练需要数千个GPU进行协作训练，涉及数十个ToR和数千个光模块和链路。在如此大规模的情况下，几乎不可能保证没有网络设备发生故障。监控和故障排除系统等工具可以在事后定位故障的根本原因，但无法防止训练崩溃。在我们的运行集群中，每个月有0.057%的NIC–ToR链接失败，并且大约有0.051%的ToR交换机遇到严重错误和崩溃。在如此高的故障率下，单个LLM训练作业每个月会遇到1~2次崩溃。此外，每天会发生5 000 ~ 60 000次链路抖动，导致模型性能的暂时下降。

2 HPN网络接入：非叠双ToR上联

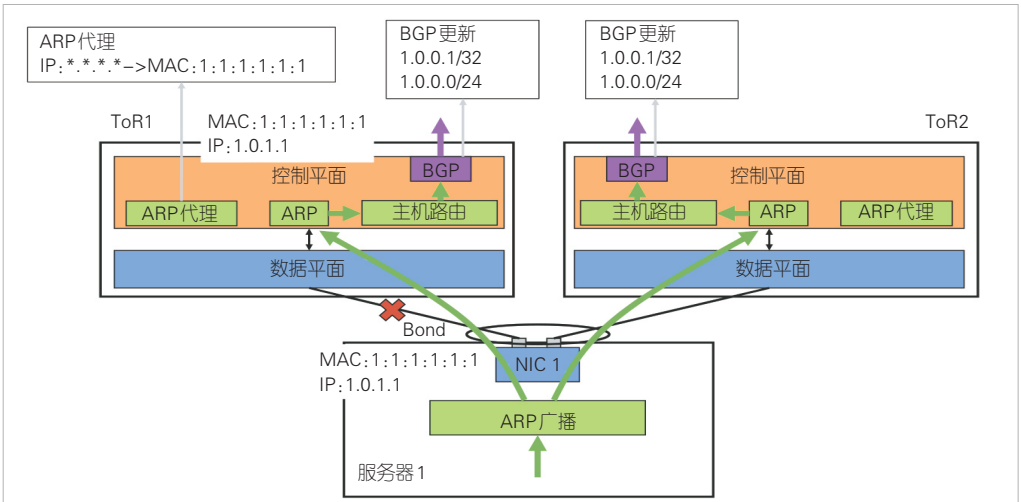
在传统数据中心网络中，每个网卡的两个端口通过一根连接到ToR交换机的电缆/光纤进行汇聚，称为单ToR设计（目前大多数云提供商广泛使用[5]）。然而，单ToR设计非常容易受到交换机/链路故障的影响，严重影响LLM训练。

非堆叠双ToR设计将每个网卡的两个端口以主–备方式连接到不同的ToR。这两个端口配置相同的IP和媒体接入控制（MAC）地址。如果一个ToR（或一个端口）宕机，另一个仍可继续工作。此外，由于同一网卡中的两个端口共享相同的队列对（QP）上下文，流量切换不会导致活动流的中断，并对上层应用透明。

然而，这样的设计引入了一个新的挑战：如何在没有直接连接的情况下同步两个不同的ToR的状态？应对这个挑战并不容易。在已有的堆叠双ToR方案中，由于两个ToR通过一个链接直接连接，它们可以通过直接链接协商一个共享的sysID。这使得主机可以通过链路聚合控制协议（LACP）与叠加式双ToR交换机进行通信。然而，因为我们想要消除ToR之间的直接链接，使它们相互独立，这意味着它们不能再使用LACP进行协商。因此，我们需要设计一种新的技术，通过一种隐式方法来“伪装”两个ToR，使主机可以通过LACP与双ToR进行通信。

如图3所示，构建非堆叠双ToR并不容易，因为我们必须确保在LACP协商过程中，双ToR交换机使用相同的MAC地址和不同的portID。我们与交换机供应商深度合作，实现了定制的LACP模块，以实现这一目标。

主机能够通过将每个ARP消息复制到NIC上的两个端口的方法来同时更新两个ToR上的ARP信息。到目前为止，所有主流的主机和交换机都能支持该非堆叠双ToR方案。



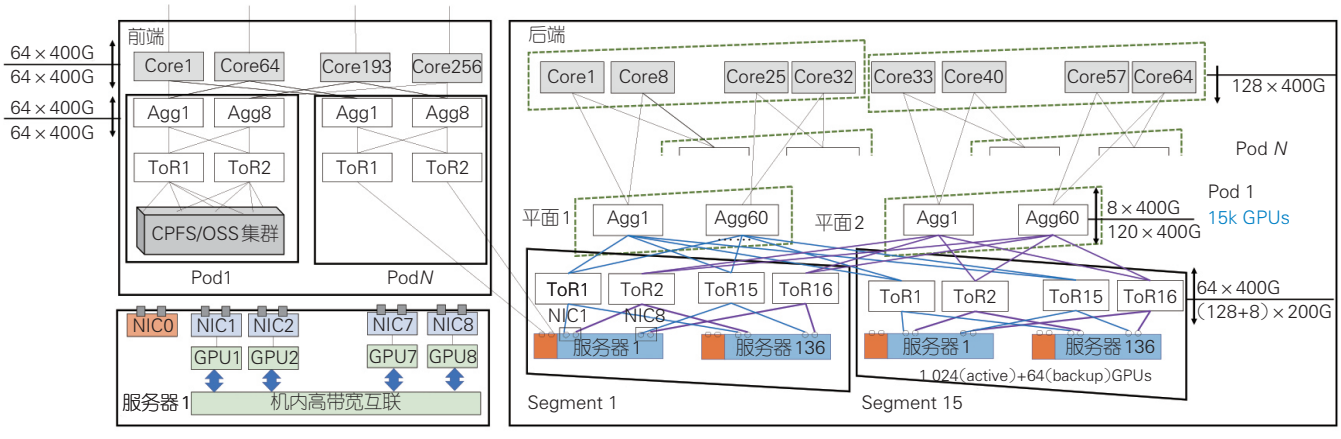
▲图3 非堆叠双上联

3 构建容纳千卡规模的一层网络

如图4所示，我们在高性能网络（HPN）中使用了最新的51.2 Tbit/s以太网单芯片交换机。在Tier1（一个Segment）中，每个交换机具有128个可用加8个备用的200 Gbit/s下行端口和60个上行的400 Gbit/s端口。这种设计确保了接近1:1的超额预订（实际上是1.067:1）。每个ToR交换机保留8个备用下行端口。我们使用这些端口连接备用主机，可以在主机端故障时快速更换主机。

1）单芯片交换机。ToR交换机的带宽容量直接决定了同一Tier1网络中GPU的数量。业界已经有支持更高带宽容量的多芯片框式交换机[6]。然而，阿里云在操作数据中心网络方面的长期经验表明，和单芯片交换机相比，多芯片框式交换机引入了更多的稳定性风险。具体来说，我们线上实际运营的单芯片交换机数量是多芯片交换机的32.6倍。相反，多芯片交换机遇到的关键硬件故障总数比单芯片交换机高3.77倍。根本原因在于多芯片交换机是一个分布式的交换系统。内部结构、芯片间相互作用、芯片与CPU的通信故障都会导致整体关键故障。因此，我们决定对所有新设计的网络架构都采用单芯片交换机。

2）多轨组网。主机内的8个GPU通过高带宽的主机内网络进行连接。虽然不同类型的GPU的主机内网络带宽不同，但是它比NIC提供的2×200 Gbit/s带宽高出4~9倍。NVIDIA是第一个提出多轨组网设计的[7]，此种网络设计已经广泛应用于训练集群中。在多轨组网中，同一铁路中的



▲图4 高性能网络整体概览

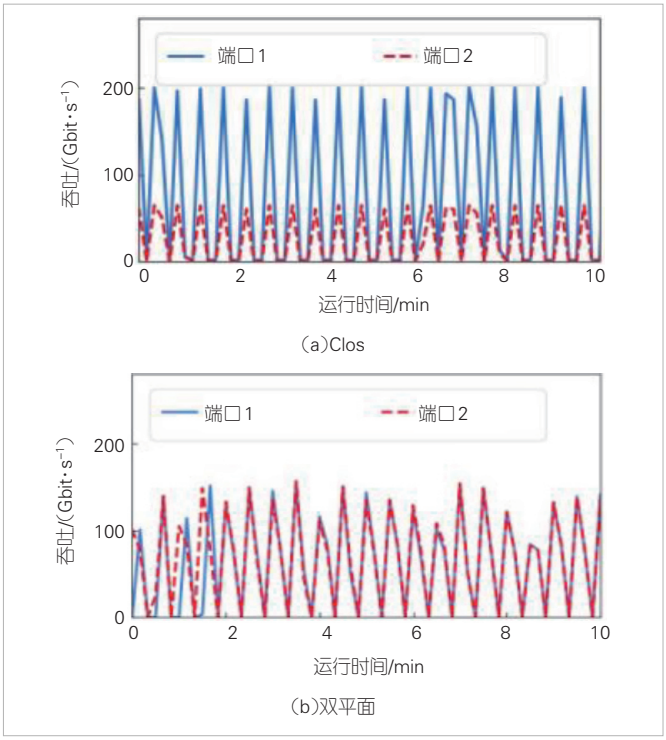
NIC通过同一套非堆叠交换机进行连接。不同轨道的NIC可以通过主机内+主机间转发的组合进行通信。例如，在图5中，如果服务器1中的GPU1想要与服务器3中的GPU2进行通信，转发路径为：服务器1的GPU1→服务器1的GPU2→ToR3→服务器3的GPU2。

4 构建容纳万卡规模的二层网络

在Tier1网络中使用双ToR，在ToR和聚合交换机之间简单部署典型的Clos拓扑结构，仍会存在哈希偏极化。在下行方向，双ToR设计导致存在2个可达下一跳，这引起了从60个聚合交换机到2个ToR交换机的高度收敛的流量。图6（a）展示了双ToR设置中两个下行端口的出口流量，流向同一网卡。我们对在生产环境中运行的GPT-3 175B的实际训练作业期间进行了测量。这两个端口的负载显著不同（吞吐量的差别高达3倍）会降低训练性能。

为了避免负载极端不均问题，我们需要在一个Pod中消除哈希偏极化。如图6（b）所示，在双平面设计中，每个双ToR设置中的ToR交换机被分为两个独立的组。有了这个设计，一旦一个流进入ToR中的任何一个上行链路，其在Pod内的转发路径就完全确定了。因此，在Pod中，哈希偏极化被完全消除了。部署双平面设计后，如图6（b）所示，不同端口的输入流量变得更加均匀，而在ToR下行端口的队列长度减少了91.8%。实际测试表明，双平面设计为跨段流量贡献了高达71.6%的性能优化。通过对512个GPU同时运行4个AllReduce作业的测试，这种优化的路径选择可以将集体通信性能提升34.7%。

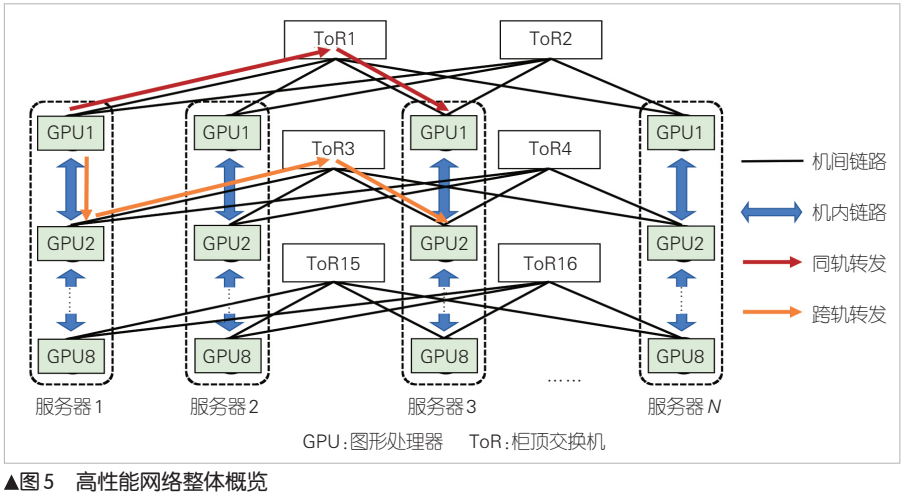
由于双平面设计，当搜索不相交路径时，我们只需搜索每个ToR交换机中的链路（即最多搜索60条链路），这样大大减少了时间消耗。HPN能够减少1或2个数量级的计算复杂性。



▲图6　同一网卡的两个端口流量

更重要的是，当发生故障时，主机只需要从ToR交换机那里获得新的等价多路径（ECMP）组，并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的ECMP组）。

双平面设计带来了另一个重要的好处：在ToR和聚合之间减少了一半的链路连接。这使得聚合交换机可以支持同一Pod中更多的Segment。因此，Tier2网络的规模翻了一番。另外，我们设置了聚合–核心的收敛比为15:1，并额外增加了聚合交换机上87.5%的端口，用于容纳更多的Segment。最终，我们实现了将15 000卡放置在同一Pod中，并为每个GPU提供了400 Gbit/s的网络接入能力。



▲图5　高性能网络整体概览

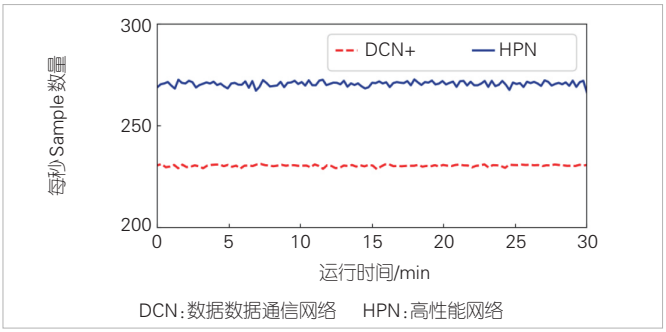
5 HPN性能评价

我们通过阿里云自主研发的大模型在集群上的训练效果来充分展示HPN所带来的性能提升。这个模型的训练采用了2 300多个GPU（超过288台服务器）。该大模型最初是在数据通信网络（DCN+）上进行训练，然后迁移到HPN上。在DCN+中，训练任务使用了19个Segment，而在HPN中，训练任务只需要3个Segment。我们观察到，迁移后性能会显著提升。图7显示，端到端训练性能提高了14.9%以上。这种端到端的

性能提升在实际生产环境中具有很大的价值。考虑到整个训练集群的构建可能会花费数十亿美元，14.9%的性能提升则可带来显著的成本节省。聚合交换机承载跨Segment流量，其统计数据直接反映网络状态。根据图8显示，跨Segment流量平均减少了37%。较少的跨Segment流量使得网络中的拥塞大幅下降。图9展示了聚合交换机下行链路队列长度分布。在DCN+中，大流量和哈希冲突不断积累队列长度；而在HPN中，该问题在很大程度上得到了解决。

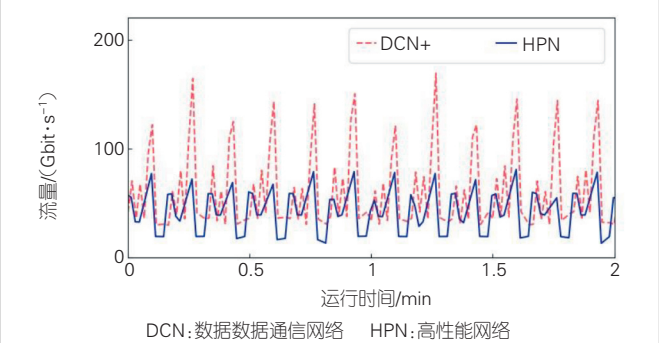
6 结束语

这篇文章介绍了HPN——一种用于大模型训练GPU集群的全新网络架构。该架构已在阿里云中大规模部署超过1年。HPN避免了传统数据中心拓扑中由单ToR设计引起的单点故障，通过双层网络连接了15 000个GPU，消除了哈希极化，并简化了最佳路径的选择。HPN使LLM训练的端到端性能提升超过14.9%。



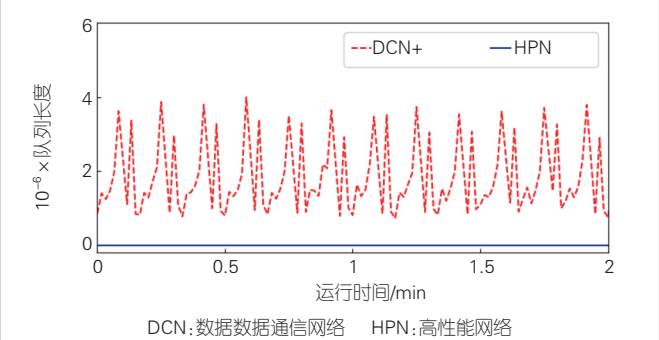
DCN：数据数据通信网络　HPN：高性能网络

▲图7　端到端训练性能



DCN：数据数据通信网络　HPN：高性能网络

▲图8　聚合层交换机入方向流量



DCN：数据数据通信网络　HPN：高性能网络

▲图9　聚合层交换机队列长度

参考文献

[1] Megatron-LM. Megatron-LM & megatron-core [EB/OL]. [2024-10-04]. https://github.com/NVIDIA/Megatron-LM

[2] DeepSpeed. DeepSpeed-extreme speed and scale for DL training and inference [EB/OL]. [2024-10-04]. https://www.microsoft.com/en-us/research/project/deepspeed/

[3] AL-FARES M, LOUKISSAS A, VAHDAT A. A scalable, commodity data center network architecture [C]//Proceedings of the ACM SIGCOMM 2008 Conference on Data Communication. ACM, 2008: 63-74. DOI: 10.1145/1402958.1402967

[4] ZHANG Z, ZHENG H, HU J, et al. Hashing linearity enables relative path control in data centers [C]// 2021 USENIX Annual Technical Conference (USENIX ATC 21). USENIX, 2021: 855-862

[5] POUTIEVSKI L, MASHAYEKHI O, ONG J, et al. Jupiter evolving: Transforming google's datacenter network via optical circuit switches and software-defined networking [C]//ACM SIGCOMM 2022 Conference (SIGCOMM '22). ACM, 2022: 66 - 85. DOI: 10.1145/ 3544216.3544265

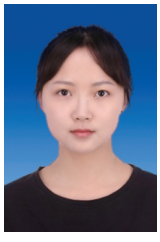
[6] Cisco. Cisco nexus 9800 series switches data sheet data sheet [EB/OL]. [2024-10-04]. https://www.cisco.com/c/en/us/products/collateral/switches/nexus-9000-series-switches/nexus9800-series-switches-ds.html

[7] FS. Troubleshoot the switchport packet loss [EB/OL]. [2024-10-06]. https://img-en.fs.com/file/user_manual/switch-port-packet-loss-troubleshooting.pdf

作　者　简　介

钱坤，阿里云高级技术专家；研究领域主要包括高性能智算网络、高性能存储网络和跨集群传输网络中的性能和稳定性优化，负责阿里云智算集群网络的监控和稳定性系统建设；发表论文10余篇。

翟恩南，阿里云资深技术专家、网络研究团队负责人，并担任SIGCOMM、NSDI、ACM SoCC等国际顶级会议程序委员会委员；研究领域包括计算机网络、分布式系统安全、程序验证等；发表论文30余篇。



操佳敏，阿里云技术专家；主要研究方向为高性能网络系统，包括面向大模型的网络性能优化、可编程芯片和可编程网络、软件定义网络等；曾参与多项国家自然科学基金、国家重点研发计划等项目；获得SIGCOMM2024最佳论文提名奖和ICCCN 2019最佳论文奖；发表论文20余篇，申请发明专利7项。

新型网络芯片技术

New Network Chip Technology

成伟/CHENG Wei，王俊杰/WANG Junjie，
杨勇涛/YANG Yongtao

（苏州盛科通信股份有限公司，中国 苏州 215125）
(Suzhou Centec Communications Co., Ltd. Suzhou 215125, China)

DOI：10.12142/ZTETJ.202406011
网络出版地址：http://kns.cnki.net/kcms/detail/34.1228.TN.20250123.1445.001.html
网络出版日期：2025-01-23
收稿日期：2024-10-17

摘要：针对大规模AI模型训练的高强度通信需求，从高性能交换架构、高性能端口、低时延、无损流控和多维负载均衡等关键技术维度，提出新型网络芯片的整体解决方案。通过整合先进网络架构与多种优化手段，该方案能够有效降低通信开销，提升训练效率，推动AI和高性能计算的规模化应用落地。

关键词：网络芯片；高性能交换架构；高性能端口；低时延；无损流控；负载均衡

Abstract: Aiming at the high-intensity communication requirements for large-scale AI model training, an overall solution of new network chip is proposed from the key technical dimensions of high-performance switching architecture, high-performance ports, low-latency, lossless flow control, and multi-dimensional load balancing. By integrating advanced network architecture and various optimization means, this solution can effectively reduce communication overhead, improve training efficiency, and promote the large-scale application of AI and high-performance computing.

Keywords: network chip; high-performance switching architecture; high-performance port; low latency; lossless flow control; load balancing

引用格式：成伟, 王俊杰, 杨勇涛. 新型网络芯片技术 [J]. 中兴通讯技术, 2024, 30(6): 68-73. DOI: 10.12142/ZTETJ.202406011
Citation：CHENG W, WANG J J, YANG Y T. New network chip technology [J]. ZTE technology journal, 2024, 30(6): 68-73. DOI: 10.12142/ZTETJ.202406011

1 新型网络芯片产业现状

随着ChatGPT等生成式人工智能（AI）的爆发式发展，AI大模型的参数规模从百亿、千亿到超万亿量级增长，这对算力资源提出了空前的需求。在Scaling law原则下，模型训练使用的算力卡数量也从万卡级别向十万卡、百万卡发展。与之对应，智算网络规模也需要同步扩大，以支持更大规模的高速无损互联。

当前，大规模智算网络互联面临两个主要挑战：一是网络设备单点带宽容量需要大幅提升，从400G、800G、1.6T向更高性能演进；二是组网更大规模演进，支持万卡、十万卡集群互联，确保端到端通信的可靠性。

高性能数据中心网络（DCN）已经通过采用Leaf-Spine（叶脊全互联架构），实现了机间网络的扩展，提高了网络的扩展性和可靠性。然而，DCN在性能优化方面仍存在以下不足：

1）带宽利用率不高。由于负载均衡和流量调度的局限，网络资源未得到充分利用。

2）时延不可控。动态负载下的拥塞和排队时延增加，影响网络的可预测性。

3）无损传输难以实现。传统传输控制协议/互联网协议（TCP/IP）难以避免丢包和重传，在高性能计算中成为瓶颈。

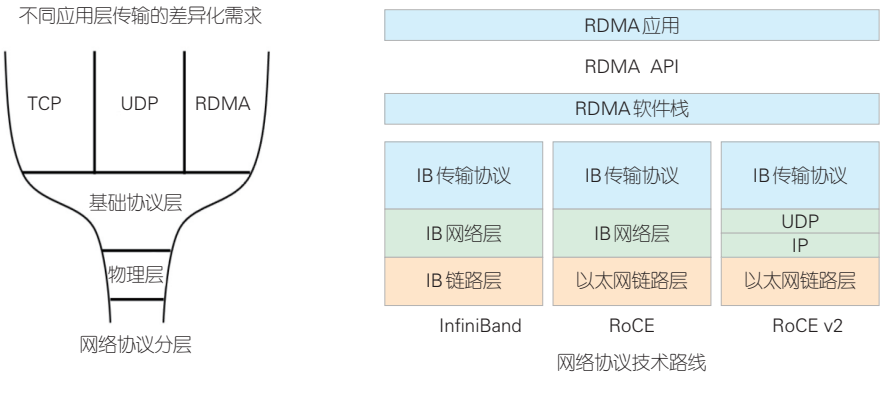
如图1所示，InfiniBand具有高带宽、低时延的特点[1-2]，但在实践中仍面临以下挑战：

1）成本高昂。专用的硬件和协议增加了部署和维护成本。

2）生态封闭。与以太网等主流网络技术兼容性差，这限制了其应用范围。

3）扩展性受限。在大规模系统中，InfiniBand网络的架构复杂度和成本成倍增加。

从机内到机间的互联扩展面临性能损失与复杂度的挑战。机内互联协议是针对芯片或单机内部的高带宽、低时延需求进行优化的，当其扩展到多节点甚至跨集群网络时，不同层级的网络通信需要不同协议转换与适配。此过程不仅引入额外的时延和系统复杂度，还造成资源利用率的下降，难



▲图1 网络协议分层与技术路线

以满足高性能计算对低时延和高效互联的要求。

1）规模扩展性。片上网络（NOC）扩展性差，难以适应大规模计算集群。网络部署方案对路由算法和拓扑设计提出了更高的要求。高速串行计算机扩展总线标准（PCIe）的拓扑结构主要为树状或点对点连接，当节点数量增加时，会出现带宽瓶颈和路由复杂度增加的问题。

2）协议兼容与转换。机内互联技术通常使用专有协议（如NVLink），而机间通信则依赖于融合以太网协议（如RoCE）。从机内互联扩展到机间互联时，必须进行协议转换和适配。这不仅增加了系统复杂度，还会引入显著的时延开销，影响高性能计算的整体性能。

2 新型网络芯片关键技术

新型网络芯片的关键技术包括高性能交换架构、高效能物理层、无损级低时延、双向联合流控、多维负载均衡、开放生态底座。网络芯片技术的发展，只有通过整合先进架构与多种网络优化技术，才能有效应对未来高性能计算和AI训练中的通信瓶颈与传输挑战。

1）高性能芯片架构。高性能芯片架构是网络交换芯片的核心。通过采用高性能的交换架构设计，网络芯片可以实现高吞吐量和低时延的数据包处理能力，满足大规模并行计算对高速数据交换的需求。

2）高性能端口。高性能端口是实现高速数据传输的关键。高速SerDes（串行器/解串器）和四电平脉冲幅度调制（PAM4）高效调制技术，使得网络芯片具备更高的物理传输速率、更低的传输损耗。

3）低时延。低时延是高性能网络的核心指标。通过简化数据处理流程，降低通信开销，尤其是在网络丢包的情况下，网络芯片能够配合流控保障无损不丢包。这样可以显著降低传输时延和业务整体时延，满足AI和高性能计算对低时延和无损的要求。

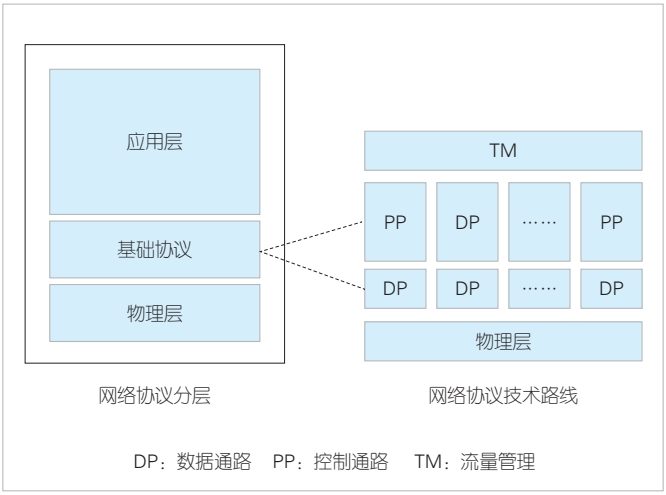
4）无损流控。无损流控是通过基于优先级的流量控制（PFC）在发送端和接收端之间建立协同机制，并根据不同的业务对数据流实施精细化的水线配置。这样可以有效防止网络拥塞和丢包，确保网络的可靠性和稳定性。

5）多维负载均衡。通过感知网络的多维状态，结合动态负载均衡算法，系统可以将本地路径决策提升到全局维度的智能调度和路径选择。这样可以有效避免网络热点和单点故障，提升整体网络吞吐量和资源利用率。

6）开放生态底座。通过支持标准化接口规范和开放标准协议，达到异构系统互联互通3个目标：通过标准协议确保系统互操作性、基于统一开放接口降低开发门槛、基于可扩展架构支持功能演进。

2.1 高性能交换架构

高性能交换架构是满足未来AI算力中心发展的基础。其中，网络芯片带宽的持续提升是必不可少的一环。如图2所示，高性能大带宽网络芯片采用多核架构，通过控制通路（PP）和数据通路（DP）按需分配来提升架构灵活性。相比于单核架构，多核架构不仅增加了前端设计的复杂度，还因



▲图2 网络协议分层与网络芯片架构

PP和DP的交互带来额外的设计开销，同时也给芯片后端布局带来多种挑战。

高性能网络芯片主要组成部分为：物理层、数据通路（DP）+控制通路（PP）+流量管理（TM）、芯片内存Memory。

1）物理层。物理层主要包含端口Serdes串行器/解串器、物理编码子层（PCS）、多媒体访问控制（MAC）等模块，这3个模块决定了芯片对外体现的端口形态和数量。

2）DP。DP是报文接收发送的物理通道，在一定程度上决定了芯片内部带宽性能。

3）PP。PP主要包含入方向PP和出方向PP。报文进到DP并在接收到一定长度后，生成一个Message，再被送往入方向PP进行处理；报文载荷是会继续接收并存储在内部缓存。

4）TM。TM主要是管理报文在Buffer中的存储和读取、队列管理。

5）芯片内存。芯片内存是指业务化的、查表的内存。根据介质不同，一般分为静态随机存取存储器（SRAM）和三态内容寻址存储器（TCAM）。TCAM常见的是访问控制列表（ACL）、掩码路由表。静态随机存取存储器主要用于二层桥接转发表、主机路由表、下一跳编辑表。

在高性能网络芯片架构设计中，性能、功耗和面积是必须考虑的3个核心指标。为实现性能、功耗、面积（PPA）的最佳平衡，需要针对具体应用场景进行权衡，以满足实际应用需求。

1）性能。通过优化芯片架构和数据通路，提升数据传输效率。

2）功耗。采用低功耗设计策略，如电源管理、时钟门控等来降低能耗。

3）面积。利用先进工艺制程和高密度集成技术，在控制芯片面积的同时增加功能模块。

4）工艺。采用更先进的半导体工艺，以降低功耗和芯片面积。

5）模块化。优化功能模块化设计布局，以提高芯片面积利用率。

2.2 高性能端口

高性能端口主要负责传输和接收数据，它将数据链路层的相关报文进行封装/解封装，在数据包之间添加/删除间隔（IPG）和起始定界符，并对传输的数据帧进行编/解码。根据对应的端口速率、传输介质类型，我们将数据转换为电信号或光信号，并通过介质发送/接收对端。

随着单芯片交换容量的提高，单端口转发能力也在不断提高。网络芯片具备全面支持400 Gbit/s端口的同时，还支持800 Gbit/s、1.6 Tbit/s的高性能端能力，这对芯片端口物理层设计提出了新的挑战。

高性能以太网端口的标准经历了多次演进，2010年的IEEE 802.3ba定义了100 Gbit/s端口；2017年的IEEE 802.3bs定义了200 Gbit/s和400 Gbit/s端口；2024年的IEEE 802.3df定义了800 Gbit/s和1.6 Tbit/s端口[3–4]。根据以太网“摩尔定律”，端口速率平均每2～3年翻一番。

如图3所示，芯片端口物理层包括物理编码子层（PCS）、物理介质连接层（PMA）、物理介质相关层（PMD）。

1）PCS。PCS主要对数据进行编/解码，并对结果进行校验。PCS位于MAC层和PMA层之间，将MAC送来的数据进行物理层编码后再送给PMA，再将PMA送到PCS的数据进行解码后再发送给MAC。

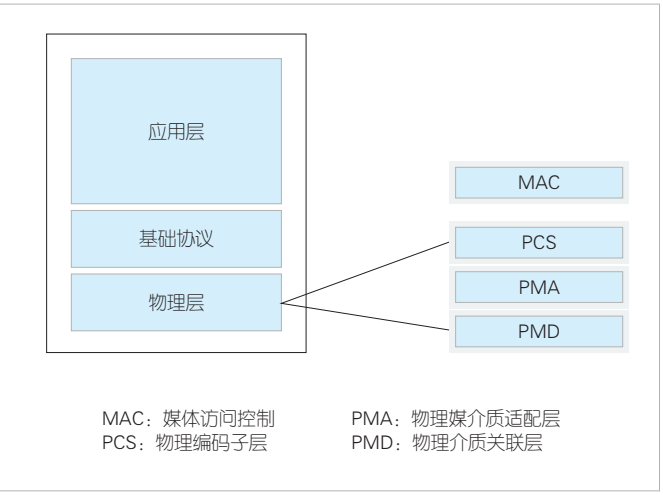
2）PMA主要用于电路的串/并转换。PMA子层集成了SerDes串行器/解串器、发送和接收缓冲、时钟发生器及时钟恢复电路。

3）PMD主要负责将串行信号转换到相应的物理介质上。物理层的PCS子层通过介质无关接口连接数据链路层，并通过PMD对外连接物理介质。

2.3 低时延

低时延是高性能网络的核心指标。实现低时延，不仅需要从芯片设计、简化业务模型处理流程等方面入手，更关键的是在网络丢包的情况下，通过网络芯片硬件级的检测和重传错误数据包。这样可以避免重新发送整个数据包，减少通信时间。

面向AI新型网络的低时延，不仅是指在网络轻负载情况下的单包测试时延，还指动态负载的实际时延，即数据流



▲图3 网络芯片与高性能端口

的完成时间[5]。如图4所示，从网络芯片时延优化的视角来看，可以将网络设备转发整体时延分解为静态时延和动态时延。

静态时延包括数据串行时延、设备转发时延和光电传输时延。这类时延由网络芯片的转发能力和链路传输的距离决定，具有确定的量级。芯片队列缓存和丢包重传对网络时延的影响是动态不可控的。

如图5所示，网络芯片时延的影响因素为芯片主频、数据包长、直通转发、端口形态、业务配置模型、实际流量模型、模块串行设计、模块并行设计。

芯片主频是直接影响芯片带宽的因素，芯片主频越高，转发带宽越大，时延就越低。芯片主频受限于工艺，随着先进工艺的提升，芯片主频也在刷新，但并不是线性的提升。同时，昂贵的流片费用也不断攀高。

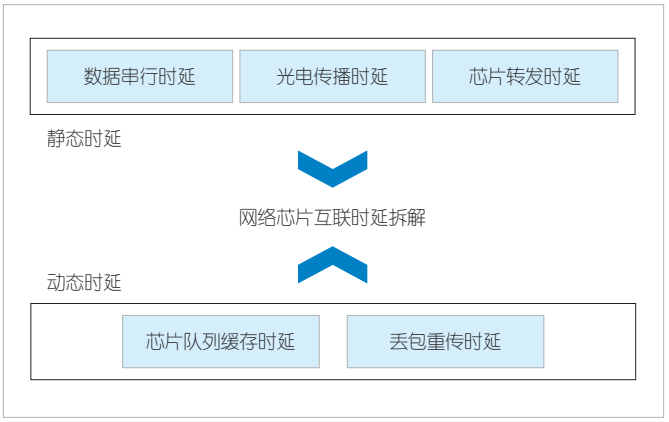
在同等条件下，相比于100G到100G端口的转发时延，全端口400G端口的转发时延会有所降低。不同业务的流量模式也会导致时延差异，64字节小包长的转发时延相比4 096字节长包的转发时延更低。

直通转发是指在数据包完整接收前即可进行后续处理。如图6所示，数据包从数据接收处理引擎直通至数据存储转发控制模块。当接收到预设数据长度（如128字节）时，系统产生信号并将数据送入下一模块处理，同时启动芯片查表转发。此时，当前数据包的后续部分持续存入数据存储转发控制模块的存储器中，并按预设长度分段送入下一级模块处理。

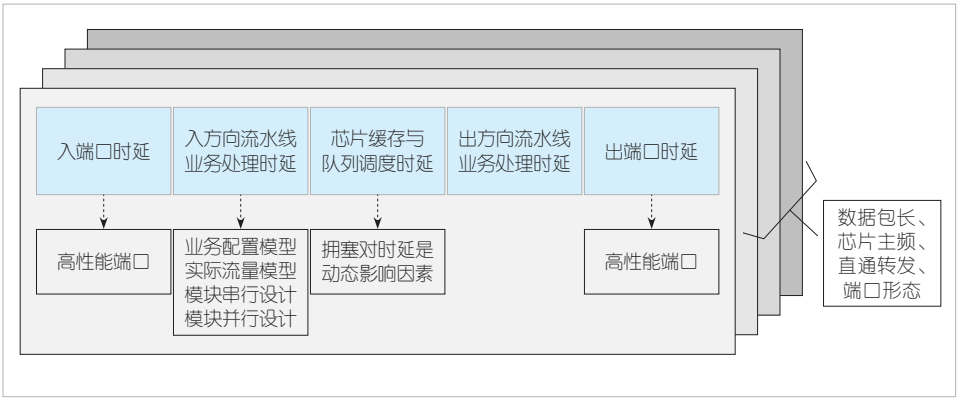
2.4 无损流控

无损流控是指PFC。通过在发送端和接收端之间建立协同机制，双向联合流控能够根据不同的业务需求对数据流进行精细化水线配置。这样可以有效防止网络拥塞和丢包，确保网络的稳定性与可靠性。

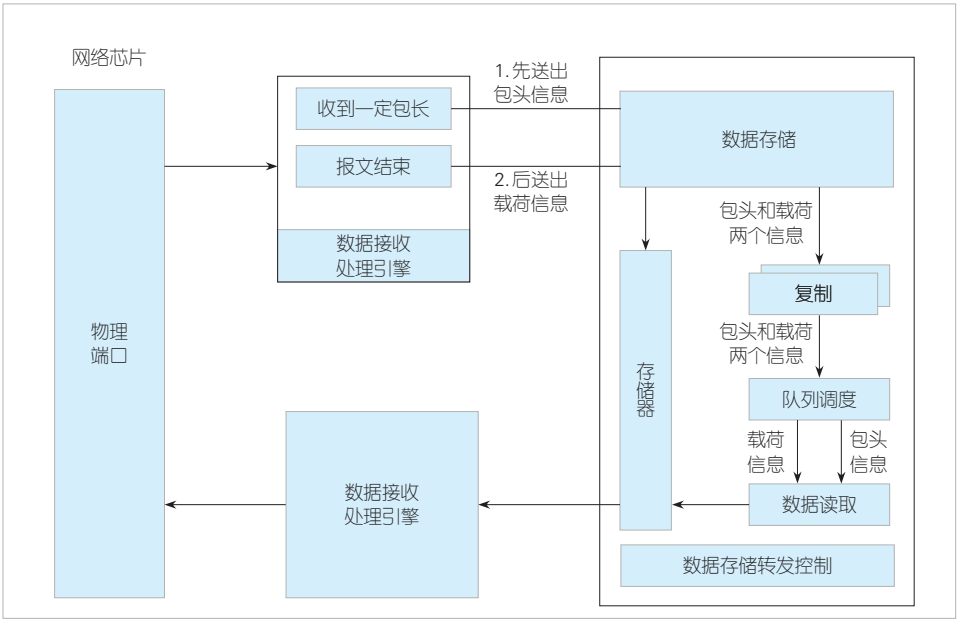
如图7所示，传统远程直接内存访问（RDMA）无损以太网采用PFC来处理拥塞场景下的丢包[5]，并使用PFC+显式拥塞通知（ECN）的方式。这在一定程度可以提前感知芯片队列拥塞，并及时通过数据中心量化拥塞通知



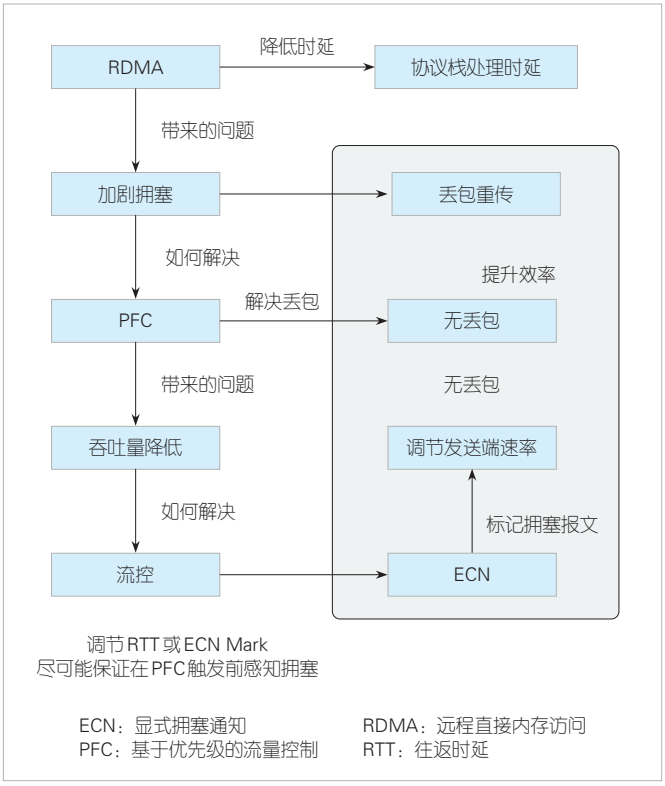
▲图4 网络互联与芯片时延



▲图5 网络芯片时延影响因子



▲图6 网络芯片直通转发模式



▲图7 远程直接内存访问无损流控机制

（DCQCN）调节发送端速率[6]，以缓解拥塞并减少丢包的出现。PFC后向流控本质上是无法解决拥塞和反馈不及时的问题。在大规模部署时，运维团队还面临PFC死锁的风险，过多的PFC Pause会降低吞吐量，同时水线的配置和调整也会给运维带来挑战[7]。

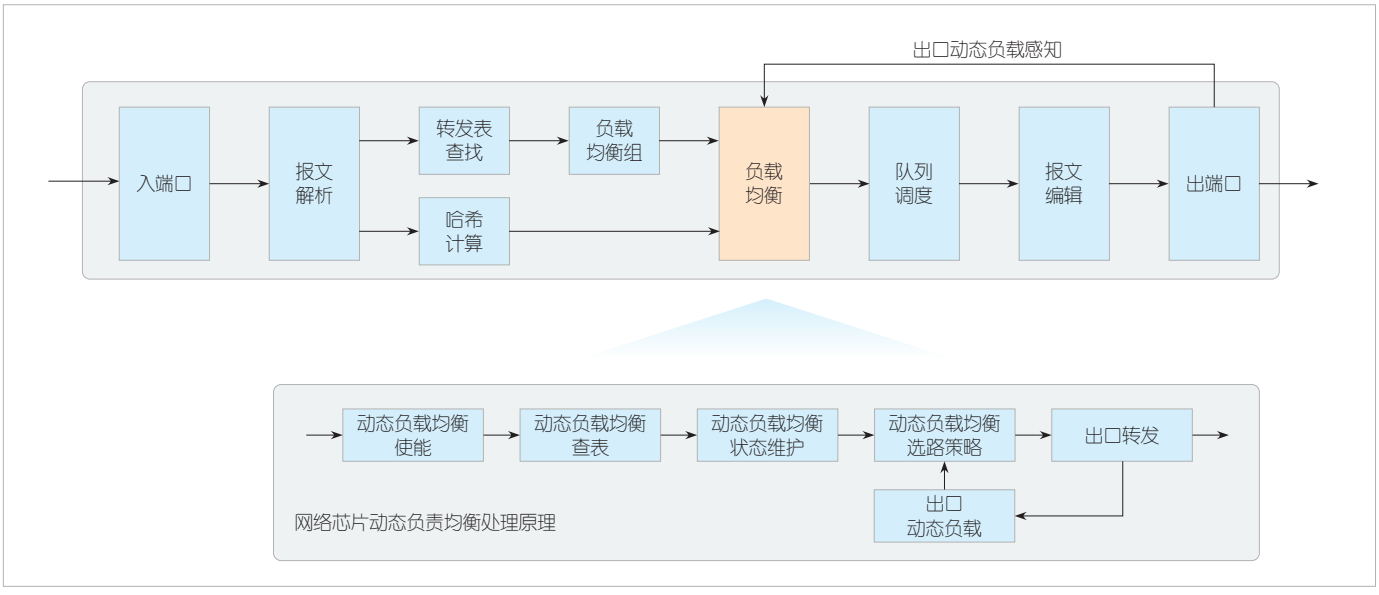
2.5 多维负载均衡

多维负载均衡是指通过感知网络的多维度状态，结合动态负载均衡算法，将本地路径决策升维到智能调度和路径选择。这样可以有效避免网络热点和单点故障，提升整体网络吞吐和利用率。

传统静态负载均衡（SLB）可以根据报文哈希值分配流量，但无法动态感知网络负载状态[8]。等价多路径路由（ECMP）无法区分不同流量的包长和带宽差异，无法感知出口动态负载。例如，当发往同一机柜顶部交换机的不同网口的转发流量通过负载均衡选路到同一下一跳设备时，该设备的下行端口容易出现拥塞。单节点的本地负载均衡算法难以解决全局冲突问题，而负载均衡哈希冲突会导致流量重叠，引发负载不均衡。为解决这一问题，动态负载均衡应运而生。

考虑到设备端口利用率，动态负载均衡将不同的子流分配到不同的端口。如图8所示，芯片动态负载均衡的处理流程为：在入方向流水线转发查表以获取下一跳负载均衡组，根据数据流特征计算哈希作为流量标识，组合生成动态负载均衡流表索引，最后基于动态负载均衡流表的维护状态更新与选路信息。动态负载均衡选路机制是关键，可以支持多种选路策略并最终选择负载均衡出口。芯片监测负载均衡出口带宽/出口队列深度的动态负载，并根据阈值配置来划分质量等级。

全局负载均衡通过引入全局网络路径信息来支持更智能的流量分配。设备接收远端设备通告报文，解析之后配置芯片表项设置出口权重。本地芯片根据收到的远端质量表，结



▲图8 芯片动态负载均衡的原理

合本地的端口质量表，计算出最终的转发出口权重。全局负载均衡能够识别人工智能网络业务的远程直接内存访问流量，并进行调度优化。在网络拥塞时，它可以通过全网智能调度来保障多路径负载均衡，实现全局业务流量的高吞吐，避免网络拥塞对远程直接内存访问性能的损失。

3 结束语

总体来看，新型网络芯片的发展需要集成高性能交换架构、高性能端口、低时延、无损流控和多维负载均衡等关键技术。只有这样，才能有效应对未来高性能计算和人工智能训练中的通信瓶颈与传输挑战。

展望未来，在开放生态系统的构建上，网络芯片技术的进一步发展将依赖于产业界多方协同推进创新。网络芯片通过支持标准化的开放协议和接口，实现与主流网络技术的深度兼容，从而推动产业链的协同合作与发展。通过构建开放、协作的生态系统，网络芯片技术能够更快速地响应市场需求，推动创新，并最终为各行各业的数字化转型提供强大的网络基础设施支持。未来，随着开放生态底座的不断完善和扩展，网络芯片技术将迎来更加光明的发展前景，为构建下一代智能互联网络奠定坚实的基础。

参考文献

[1] Infiniband. Infiniband architecture specification [EB/OL]. [2024-10-13]. https://www.infinibandta.org/ibta-specification/
[2] GUO C X, WU H T, DENG Z, et al. RDMA over commodity ethernet at scale [EB/OL]. [2024-10-12]. https://dl.acm.org/doi/10.1145/2934872.2934908
[3] Ethernet Alliance. 2024 Ethernet roadmap [EB/OL]. [2024-10-10]. https://ethernetalliance.org/technology/ethernet-roadmap/
[4] IEEE. IEEE 802.3 ethernet working group [EB/OL]. [2024-10-10]. https://www.ieee802.org/3/
[5] 开放数据中心委员会. ODCC无损网络技术白皮书 [R]. 2017
[6] IEEE. The lossless network for data centers [R]. 2018
[7] 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
[8] 沈耿彪, 李清, 江勇, 等. 数据中心网络负载均衡问题研究 [J]. 软件学报, 2020, 31(7): 2221-2244
[9] 毛鹏轩. 下一代网络拥塞控制关键算法的研究 [D]. 北京: 北京交通大学, 2013

作 者 简 介

成伟，苏州盛科通信股份有限公司副总裁；负责产品、战略和产业生态，主要研究方向为高性能网络、确定性网络、边缘计算网络、可编程网络、光电融合网络领域。

王俊杰，苏州盛科通信股份有限公司标准总工；负责技术标准工作，主要研究方向为高性能互联、网络协议、开源系统等。

杨勇涛，苏州盛科通信股份有限公司资深总监；负责网络交换芯片的技术市场与推广工作，对软件定义网络与白盒交换机有深入研究。

网络协议的演进和创新

Evolution and Innovation of Network Protocols

李星/LI Xing，包丛笑/BAO Congxiao
（清华大学，中国 北京 100084）
(Tsinghua University, Beijing 100084, China)

DOI：10.12142/ZTETJ.202406012
网络出版地址：http://kns.cnki.net/kcms/detail/34.1228.TN.20250109.0925.002.html
网络出版日期：2025-01-09
收稿日期：2024-10-15

摘要：总结了传输控制协议/互联网互联协议（TCP/IP）的分布式架构、无连接传输、尽力而为的服务模型、端对端通信模型和开放性5个基本特征，指出这些特征是TCP/IP优于其他网络协议的根本原因。针对新的网络需求，分析这些特征的异化。指出在当前快速变化的技术环境中，TCP/IP作为互联网的基石，其核心思想依然具有重要意义。要充分利用互联网协议第6版（IPv6）的机遇，努力创新以适应人工智能等新技术发展带来的挑战。

关键词：网络体系结构；网络协议；TCP/IP；IPv6

Abstract: The five basic characteristics of transmission control protocal (TCP)/Internet protocal (IP) are summarized, including distributed architecture, connectionless transmission, best-effort service model, end-to-end communication model, and openness. It points out that these characteristics are the fundamental reasons TCP/IP is better than network protocols. The alienation of these features for the new network requirements is analyzed. It also points out that in the current rapidly changing technological environment, TCP/IP, as the cornerstone of the Internet, its core idea is still of great significance. It is necessary to make full use of the opportunities of IPv6 and strive to innovate to meet the needs brought about by new technologies such as artificial intelligence.

Keywords: network architecture; network protocol; TCP/IP; IPv6

引用格式：李星, 包丛笑. 网络协议的演进和创新 [J]. 中兴通讯技术, 2024, 30(6): 74-83. DOI: 10.12142/ZTETJ.202406012
Citation：LI X, BAO C X. Evolution and innovation of network protocols [J]. ZTE technology journal, 2024, 30(6): 74-83. DOI: 10.12142/ZTETJ.202406012

1 TCP/IP的发明和网络技术演进

2024年是ARPANET诞生55周年，也是中国全功能接入互联网30周年。网络协议演进过程的回顾，对把握未来网络技术的发展方向具有重要的参考意义。

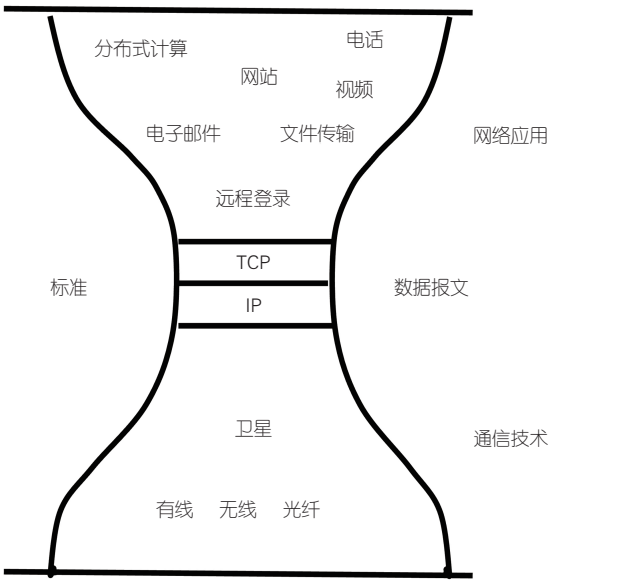
1.1 计划中的信息高速公路

人类于1852年发明了电报，1876年发明了电话，1946年发明了数字计算机（ENIAC），1957年发射了世界上第一颗人造地球卫星（Sputnik），1958年发明了集成电路。随着数字计算机、地球同步通信卫星和集成电路等技术的发展，模拟电话于20世纪70年代开始向数字电话演进。当时计划先从模拟电话网演进到综合数字网（ISDN），演进到窄带综合业务数字网（N-ISDN），并在“信息高速公路”概念的促进下，演进到宽带综合数字网（B-ISDN），最终成全球信息基础设施（GII）。虽然计划很完美，但传输控制协议/互联网协议（TCP/IP）的出现，改变了一切[1-2]。

1.2 ARPANET和TCP/IP

ARPANET是世界上第一个实现分组交换技术的计算机网络，ARPANET由美国国防部高级研究计划局（DARPA）于20世纪60年代末开始部署，旨在实现不同地理位置之间的计算机通信和资源共享，提供可靠的通信手段，促进科研合作，增强国家安全。在ARPANET的早期阶段，网络控制程序（NCP）是主要的通信协议。然而，随着网络规模的扩大和需求的增加，NCP的局限性逐渐显现出来。首先，NCP仅适用于特定类型的网络硬件，缺乏灵活性和可扩展性；其次，在处理复杂通信任务时表现不佳，难以满足日益增长的应用需求。为了突破这些局限性，研究人员开始探索新的网络协议。20世纪70年代初，人们提出了TCP/IP。TCP/IP采用分层架构，极大地提高了网络的可扩展性和兼容性。互联网体系结构的“细腰模型”如图1所示。

1983年1月1日，ARPANET正式从NCP过渡到TCP/IP，这一转变标志着互联网时代的真正开始[3]。



▲图1　互联网的“细腰模型”

1.3 TCP/IP 的基本特性

TCP/IP 协议族作为互联网的核心协议，具有多种特性。这些特性使得 TCP/IP 在全球范围内广泛应用并取得了巨大的成功[4–5]。

1）分布式架构

TCP/IP 采用了分布式设计理念。这意味着网络中的每个节点都具有相同的地位和功能，没有集中控制的核心节点。分布式架构的主要优势在于其高可靠性和灵活性。由于没有单一的控制中心，网络的故障点减少，即使某个节点出现问题，也不会影响整个网络的正常运行。

2）无连接传输

TCP/IP 中的 IP 采用了无连接传输的原理。无连接传输意味着每个数据包在网络中独立传输，不依赖于其他数据包或连接状态。这种设计使得网络能够处理大量并发数据流，提高了整体吞吐量和效率。

3）尽力而为的服务模型

TCP/IP 采用了尽力而为（Best–Effort）的服务模型。这种模型不保证数据包一定能够到达目的地，也不提供任何形式的质量保证或错误恢复机制。尽力而为地实现依赖于网络中各节点的自主决策和路由算法，每个数据包在传输过程中会根据当前网络状态选择最优路径。

4）端对端通信模型

TCP/IP 采用了端对端的通信模型。这意味着数据传输从源端到达目标端时，中间网络设备不对数据内容进行修改或处理。端对端通信模型在网络设计中具有重要意义：它简化了网络层的功能，使得应用层可以根据需要实现各种复杂功能。此外，端对端通信模型增强了数据传输的透明性和安全性，确保数据在传输过程中不会被篡改或泄露。

5）开放性

TCP/IP 协议族具有高度的开放性，这主要体现在其标准化流程上。互联网工程任务组（IETF）是负责 TCP/IP 标准化的主要机构。IETF 通过公开讨论和协作，确保所有相关方都能参与到标准制定的过程中。标准化流程通常包括以下几个步骤：提出草案、公开评审、修订并最终发布为 RFC（IETF 发布的一系列备忘录）文档。这种开放的标准化流程不仅促进了技术创新，还确保了协议的广泛接受和兼容性。开放性使得 TCP/IP 能够快速适应新技术和新需求，持续推动互联网的发展。

基于 TCP/IP 的互联网也是开放的，体现在地址的唯一性、域名的唯一性和协议标准和协议参数的唯一性上。边界网关协议（BGP）可以使各自管理独立的自治域实现互联互通。

1.4 协议之战

1）TCP/IP 与 X.25 之战

X.25 是一种早期的分组交换网络协议，由国际电信联盟（ITU）于 1976 年制定。它主要用于数据通信和远程终端连接，并在当时的公共数据网（PDN）中占据重要地位。X.25 采用面向连接的传输模式，提供可靠的错误检测和纠正机制，确保数据包能够准确到达目的地。X.25 虽然在早期占据重要地位，但其复杂性和灵活性不足使其逐渐被边缘化。TCP/IP 的成功证明了分布式、简洁、灵活和透明的协议设计在现代通信中的重要性[6]。

2）TCP/IP 与 ISDN 之战

ISDN 是一种在 20 世纪 80 年代和 90 年代业界推广的电信标准。它旨在提供高带宽的数字传输服务，能够同时处理语音、数据和视频等多种类型的通信需求。N–ISDN 协议采用了 2B+D 的结构：两个 B 信道用于传输数据，每个信道提供 64 kbit/s 的带宽；一个 D 信道用于控制信令，提供 16 kbit/s 或 64 kbit/s 的带宽。最终，TCP/IP 协议在与 ISDN 的竞争中胜出。主要原因包括：TCP/IP 的灵活性和可扩展性使其能够适应不断变化的互联网环境，而 ISDN 则显得相对僵硬。N–ISDN 一度成为互联网“最后一公里”的接入技术，称为“一线通”，但其成本和速率无法与后续的非对称数字用户线路（ADSL）和无源光纤网络（PON）技术竞争，而其话音强项又被网络电话（VoIP）完全替代[7]。

3）TCP/IP与ATM之战

异步传输模式（ATM）于20世纪80年代被业界提出，是一种高带宽、低延迟的网络技术，最初用于电信公司的语音和数据通信。ATM采用固定长度的53字节单元（称为信源）进行数据传输，这种设计使其能够提供高效的多媒体传输服务，包括实时语音、视频和数据。在20世纪90年代，ATM被认为是未来网络技术的重要组成部分，特别是在广域网（WAN）中的应用。最终，TCP/IP在与ATM的竞争中胜出，主要原因为：ATM与TCP/IP的分布式、无连接、尽力而为和端对端的特征完全是相反的，因此在大规模网络中面临可扩展性问题；虽然ATM提供了高质量服务，但其复杂性和成本在动态网络环境中无法实现。在1995年，ATM 155 Mbit/s的带宽比三次群T3（45 Mbit/s）或E3（34 Mbit/s）有优势，因此美国超高速网络服务（vBNS）使用了classic IP over ATM技术。在局域网中的ATM技术如 局域网仿真（LANE）和ATM支持多协议（MPOA）与IP over Ethernet的技术相比完全没有优势，但逐步衍生出了多协议标签交换（MPLS）[8]技术。

4）TCP/IP与私有协议之战

在互联网发展的早期阶段，TCP/IP还面临着多种私有协议的挑战。这些私有协议包括系统网络架构（SNA）、网络输入输出系统（NetBIOS）/ NetBIOS扩展用户界面（Net - BEUI）、数字设备公司网络（DECnet）、虚拟网络交换（VINES）、Xerox网络系统（XNS）、UNIX到UNIX复制（UUCP）、苹果交谈协议（AppleTalk）和互联网数据包交换（IPX）/序列分组交换协议（SPX）等。虽然它们在各自领域内占据重要地位，但最终都未能抵挡TCP/IP的崛起[9]。

（1）SNA

SNA是由IBM开发的网络架构，主要用于大型企业和政府机构的数据通信。它提供了强大的错误检测和纠正功能，适用于高可靠性需求的场景。SNA在网络层和更高层次上都采用面向连接的架构。

（2）NetBIOS/NetBEUI

NetBIOS是一种应用程序编程接口（API），该接口为开放系统互联（OSI）模型的会话层提供服务。它允许不同计算机上的应用程序通过LAN进行通信。NetBEUI是一种简单的轻量级网络协议，用于小型局域网。NetBIOS在会话层是面向连接的，但可以在网络层上运行无连接和面向连接的传输协议。

（3）DECnet

DECnet是由Digital Equipment Corporation开发的网络协议，主要用于虚拟地址扩展（VAX）和程序数据处理机（PDP）之间的通信。它提供了高效的数据传输和管理功能。DECnet IV在网络层采用面向连接的架构，在通信实体之间建立逻辑连接，以确保数据的可靠和有序传输。DECnet V（DECnet/OSI）在网络层支持面向连接和无连接通信。

（4）VINES

VINES是由Banyan Systems开发的网络协议，主要用于企业级局域网通信。它提供了高效的路由和服务定位功能。在网络层采用面向连接的架构。

（5）XNS

XNS是由施乐公司开发的网络协议，主要用于局域网通信。它提供了高效的数据传输和管理功能。XNS中的网络层协议是无连接的。

（6）AppleTalk

AppleTalk是由苹果公司开发的网络协议，主要用于Macintosh计算机之间的通信。它提供了高效的数据传输和管理功能。AppleTalk中的网络层协议是无连接的。

（7）IPX

IPX是一种网络协议，最初由Novell开发，用于其Net - Ware操作系统。它主要用于在计算机网络中传输数据包，特别是在LAN和WAN环境中。IPX中的网络层协议是无连接的。

总之，基于面向连接架构的私有网络协议与TCP/IP比较，灵活性、简单性差一些。虽然私有网络协议也有无连接的，但这些协议基本上是基于局域网的，不具备可扩展性。同时，所有私有网络协议都不具有开放性，因此在与TCP/IP的对决中失败是必然的。

5）TCP/IP与OSI之战

OSI模型是由国际标准化组织（ISO）在20世纪80年代提出的网络通信标准。OSI模型将网络通信过程分为7层，每一层具有不同的功能。这种分层设计使得各层之间独立运行，便于模块化开发和维护。OSI模型在当时被认为是网络通信标准的未来方向，具有系统性和科学性的特点。但是，OSI模型在网络层提供了面向连接的协议（如X.25）和无连接网络协议（CLNP），这两种协议分别满足不同的通信需求。OSI模型在传输层只有面向连接的TP4协议，提供可靠的数据传输服务。与之相比，TCP/IP的网络层协议IP采用无连接传输方式，数据包独立传输，不需要建立连接。TCP/IP的传输层包括TCP和用户数据报协议（UDP）。TCP提供面向连接的可靠传输，而UDP则提供无连接的不可靠传输服务。这种选择使得TCP/IP既简单适用于各种链路层协议（IP over Everything），又能够灵活地应对各种应用的需求（Everything over IP）[10]。

2 TCP/IP 网络的异化

TCP/IP 的“分布式”“无连接”“尽力而为”“端对端”和“开放性”的技术特征带来的最大优势是“可扩展性”，而 TCP/IP 的“安全性”和“服务质量”是相对的弱项。研究人员希望从体系结构上改进 TCP/IP，以使弱项变强，但却有意识或无意识地把 TCP/IP 的技术特征异化。

2.1 去“分布式”的异化

1）软件定义网络（SDN）

SDN 的主要目标是通过将网络控制与数据转发分离来简化网络管理，提高灵活性。其技术特点为：控制平面与数据平面分离，集中化管理，通过虚拟化技术支持网元的可编程性。由于 SDN 是中心式而不是分布式的架构，其所面临的问题为：控制器有单点故障风险，总体复杂性增加，兼容性差。SDN 可以在特定场景下（如数据中心、企业内部网络等）提供更高的灵活性和管理效率，而不可能在多个独立管理域的情况下被大规模使用[11]。

（1）域名系统（DNS）和域名系统安全扩展（DNSSEC）

DNS 是一个分布式数据库系统，用于将人类可读的域名（如 www.example.com）转换为计算机可理解的 IP 地址。这使得互联网上的资源更容易被访问和管理。尽管 DNS 是一个分布式系统，但其“根”服务器和顶级域名服务器在某种程度上具有中心化的特征。这可能导致单点故障、性能瓶颈和安全风险。DNSSEC 是一套用于增强 DNS 安全性的协议扩展，通过数字签名来确保 DNS 查询响应的真实性和完整性。DNSSEC 通过引入信任链和密钥管理机制，增强了对中心化服务器（如根服务器和顶级域名服务器）的依赖性。这在某种程度上加强了 DNS 系统的中心化趋势。因此，DNS 的运行确实存在着解析链断裂的问题，这给网络的可生存性带来风险。目前各种解决方案的本质是试图使域名系统重新具有“分布式”的特性[12]。

（2）路由安全和资源公钥基础设施

互联网上运行 BGP 的路由系统是典型的分布式系统，但是这一系统面临着路由劫持等安全风险。近年来互联网引入了资源公钥基础设施（rPKI）。rPKI 是一种基于密码学的安全框架，它可以通过验证 IP 地址和自治系统号码（ASN）的所有权来防止 BGP 路由劫持等问题。rPKI 系统由 5 个信任锚点构成，每个信任锚点由一个区域互联网注册管理机构（RIR）管理。这些 RIR 包括亚太网络信息中心（APNIC）、北美互联网号码分配机构（ARIN）、非洲网络信息中心（AFRINIC）、拉丁美洲网络信息中心（LACNIC）和欧洲网络协调中心（RIPE NCC）。虽然 rPKI 系统有 5 个信任锚点，但 rPKI 的运行确实存在着丢锚的问题，给网络的可生存性带来风险[13]。

2.2 去“无连接”的异化

1）OpenFlow

OpenFlow 是一种 SDN 技术，它通过将控制平面与数据平面分离，使网络设备的转发行为由中心化的控制器动态管理。OpenFlow 可以被认为是一种面向连接的技术，因为中心控制器根据 OpenFlow 五元组（源 IP 地址、目的 IP 地址、源端口号、目的端口号和协议类型）甚至更多的特征来决定数据包的转发路径。OpenFlow 交换机维护一个或多个流表，每个流表包含一组流条目。每个流条目包括匹配域、优先级、计数器、操作集合和超时值。OpenFlow 具有的特点为：灵活性、可编程性、网元的可管理性和控制接口的开放性，但是没有了无连接的特性。这使得 Openflow 的问题，例如可扩展性差、复杂度高、兼容性差等突显出来。OpenFlow 可以在特定场景下（如数据中心、企业内部网络等）提供更高的灵活性和管理效率，但却不能在多个独立管理域的情况下大规模使用[14]。

2）MPLS

MPLS 是一种高效的网络传输机制，它结合了第二层和第三层的功能。在数据包前面加上固定长度的标签进行路由和转发决策。MPLS 通过引入标签和标签交换路径（LSP），实现了一种类似于虚电路的面向连接机制。这使得 MPLS 在提供高效转发、流量工程和服务质量（QoS）保证方面具有显著优势。虽然 MPLS 在某些方面仍然保留了无连接网络的特性（如 IP 地址的路由决策），但其核心机制是基于面向连接的原则设计的。因此，我们可以认为 MPLS 是一种面向连接的技术。MPLS 一般在单个自治域内部署，无法扩展到整个互联网[15]。

3）分段路由（SR）和基于 IPv6 转发平面的 SR

SR 是一种源路由技术，它允许数据包沿着预定义的路径在网络中传输。SRv6 是将 SR 应用于 IPv6 网络的具体实现。SRv6 使用分段 ID（SID）来表示网络中的节点或路径段。每个 SID 可以表示一个 IPv6 地址或其他类型的标识符。SRv6 允许在源节点上对数据包进行路径编程，即在发送数据包时指定其完整的传输路径。这使得网络可以根据业务需求和策略动态调整数据流的路径。SRv6 在极端情况下，可以类比为面向连接的模式。SRv6 的特点为：当每一跳都被明确定义时，数据包的传输路径是确定的，从而可以更精细地控制流量分布，优化网络资源利用率；由于路径是预先定义的，SRv6 也可以降低数据包在网络中被随意转发的风险，

增强安全性。一般认为SRv6可以根据分段的多少，在面向连接和无连接之间寻找平衡，以达到预设的目标[16]。

2.3 去“尽力而为”的异化

为了确保网络QoS，避免在“尽力而为”模式下可能出现的性能问题，研究人员开发了若干技术[17]。

1）集成服务

集成服务（IntServ）的典型实现为资源预留协议（RSVP），用于预留网络资源，确保特定数据流获得所需的带宽、延迟和抖动。RSVP能够为每个数据流提供精确的QoS保证，支持动态资源预留和释放，适应网络变化。但是RSVP可扩展性差，在大规模网络中，管理大量数据流的资源预留非常复杂，且资源消耗大，需要在每个节点上维护大量状态信息，这增加了网络设备的负担。

2）区别服务（DiffServ）

DiffServ通过在数据包头部添加流类标记DSCP字段来标识不同的服务类别，比InterServ有更好的可扩展性。DiffServ简化了网络中的QoS管理，但是难以为单个数据流提供精确的QoS保证。同时，DiffServ需要在整个网络中统一配置和管理QoS策略，否则可能出现不一致性。

2.4 去“端对端”的异化

传统的TCP/IP是端对端的。动态地址分配、地址转换设备（NAT）和内容分发网络（CDN）的引入，使互联网不再支持端对端的特性[18]。

1）地址转换设备

由于IPv4的地址耗尽问题，地址转换设备（NAT）技术广泛应用于客户机联网。NAT将内部私有IP地址映射到外部公共IP地址，以便在公共网络中进行通信。为了区分多个内部主机的流量，NAT设备会使用不同的端口号进行映射。NAT的特点是：缓解IPv4地址短缺问题，能够在对上游网络运营商的情况下有效地进行流量调度，简化了网络管理。NAT破坏了端对端连接，带来了连接穿透问题，并因为协议兼容性增加了溯源的难度。

2）CDN

CDN通过在全球范围内部署大量的服务器节点，将内容缓存到距离用户最近的节点上，以减少延迟，提高访问速度。CDN根据网络状况、服务器负载和用户位置等因素动态分配流量，确保资源利用率最大化；通过缓存静态内容（如图片、视频、层叠样式表文件等）和部分动态内容，减少源站的负载并加快响应速度；通过DNS解析技术将用户请求重定向到最佳的服务器节点，实现智能路由和流量优化；提供多层次的安全防护，包括分布式拒绝服务（DDoS）攻击防御、安全套接层（SSL）/传输层安全性协议（TLS）加密、Web应用防火墙（WAF）等，确保内容传输的安全性。但是CDN使用分布式服务器节点，用户请求被重定向到最佳的节点上，源站的IP地址并不唯一。这打破了传统TCP/IP模型中端对端通信的直接性，使得数据传输路径变得更加复杂，难以进行端到端的追踪和监控。CDN也存在缓存一致性问题。在动态内容频繁更新的场景下，确保所有节点上的缓存内容一致性对CDN来说是一个挑战。

2.5 去“开放性”的异化

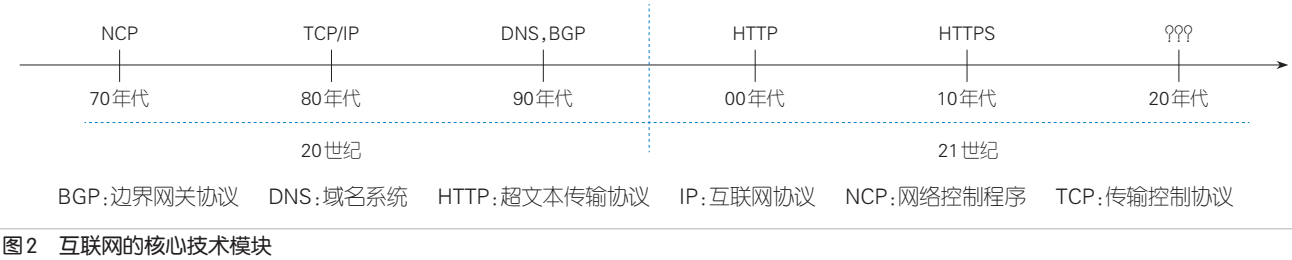
总体来说，目前互联网标准的制定过程依然延续着IETF的工作流程，保持了TCP/IP标准的开放性。但基于某些原因，若干设备厂商和运营商也在推行不经过IETF流程的私有协议。在专网和“围墙花园”的场景下，网内确实可以采用未经IETF标准化的私有协议，也不需要全球的地址唯一性和域名唯一性，但这确实是一种对于“开放性”的异化，从长远的视角看会对全球互联网的互联互通带来危害[19]。

3 互联网的核心技术模块

以10年为一个周期，过去50年来最具代表性的互联网核心技术模块如图2所示。

3.1 NCP

20世纪70年代业界的核心技术是网络控制程序（NCP）



▲图2 互联网的核心技术模块

协议。NCP是ARPANET在20世纪70年代的核心技术组件，是世界上第一个规模部署的分组网络，为主机间通信管理、数据传输和流量控制提供了基本功能。值得一提的是NCP协议是一个面向连接的协议。尽管NCP最终被TCP/IP取代，但它在互联网早期的发展中扮演了关键角色，为现代网络协议的设计和实现积累了宝贵经验[3]。

3.2 TCP/IP

20世纪80年代业界的核心技术是TCP/IP。TCP/IP是一组通信协议，用于在互联网上的数据传输。这些协议由DARPA资助开发，最初用于ARPANET的改进和扩展，之后IETF接手了其标准化方面的工作。TCP/IP协议是现代互联网的基础，提供了可靠、灵活和高效的通信能力。TCP/IP协议栈在20世纪80年代的推广和标准化为互联网奠定了基础，其分层架构、可靠传输机制和灵活的路由算法使得全球范围内的计算机网络能够高效地进行通信和数据交换[3]。

TCP/IP中的IP是网络层协议，TCP是传输层协议。

1）IP

IP实现了TCP/IP的“分布式”“无连接”和“尽力而为”的技术特征。目前仍然广泛使用的是版本4（IPv4），但其面临着地址耗尽的问题。IPv6是下一代的网络层协议。全球互联网目前仍然处于从IPv4到IPv6的过渡阶段。

2）TCP等

TCP等传输层协议实现了TCP/IP的“端对端”技术特征。在早期的互联网设计中，TCP和IP是紧密结合在一起的。然而，随着需求的增加和应用场景的多样化，两者的分离成为必要。这主要有两个原因：一是，不同的应用对传输层协议有不同的需求，例如，TCP提供可靠的数据传输，而UDP则提供无连接、尽力而为的服务，适合于实时性要求高但可以容忍一定丢包率的应用；二是，分离后的IP层能够更好地处理路由和网络层的问题，这使得整个网络架构更加灵活和模块化。这种分离不仅简化了协议设计，还为后续的扩展和优化提供了便利。

TCP的拥塞控制算法是TCP成功的关键因素之一，包括慢启动、指数增长和后退机制。通过这些机制，TCP能够有效地避免网络拥塞，提高整体性能。在慢启动阶段，发送端逐步增加发送窗口大小，直至检测到丢包；在指数增长阶段，迅速扩大传输速率，以充分利用可用带宽；后退机制在检测到拥塞时减少发送窗口，从而缓解网络负担。这些算法共同作用，确保了数据传输的高效性和稳定性，为互联网的顺利运行提供了坚实的基础。

从广义上来讲，任何在网络层之上的协议都可以被视为传输层协议，具体由IPv4报头中的“协议”或IPv6的“下一个报头”来定义。这意味着可以有256种广义传输层协议，包括但不限于TCP和UDP。但是，互联网严格意义上的传输层协议只有TCP和UDP取得了广泛的应用，其他协议如数据报拥塞控制协议（DCCP）、流控制传输协议（SCTP）等并未获得同样的成功，原因主要在于这些协议在设计和实现上存的局限性。此外，TCP和UDP已经形成了庞大的生态系统和标准化体系，新协议难以替代它们。

运行在TCP之上的TLS取得了巨大的成功。TLS通过加密数据传输和身份验证，有效地防止了数据窃听、篡改和伪造等威胁。运行在TLS之上最典型的应用层协议是超文本传输协议（HTTPS），它广泛应用在电子商务、金融交易和社交媒体等领域，确保了数据传输的安全性和隐私保护。

在UDP基础上运行的快速UDP互联网连接（QUIC）协议集成了端对端的数据一致性、安全性和流控，是传输层协议的重大进展。QUIC通过减少握手时间和改进拥塞控制算法，显著提高了数据传输的效率和可靠性。其多路复用机制还能够在单个连接上并行传输多个流，从而优化资源利用，降低延迟。

未来，传输层协议的发展可能包括以下几个方向：进一步优化拥塞控制算法，以适应更加复杂和动态的网络环境；增强安全性和隐私保护措施，以应对日益严峻的网络威胁；探索基于人工智能和机器学习的自适应传输策略，以提高数据传输的智能化水平。

3.3 BGP和DNS

20世纪90年代业界的核心技术是BGP和DNS。

1）BGP

BGP是一种在自治系统之间交换路由信息的互联网协议。作为外部网关协议（EGP）的一部分被引入，BGP具有优越的可扩展性，最终取代了EGP。1994年，BGP版本4（BGP–4）在RFC 1654中正式发布，这是目前广泛使用的版本。BGP–4引入了多个重要特性：支持路径向量协议，拥有自治域号码（ASN）和路径属性，支持策略控制、循环检测，具备扩展性。BGP–4的引入极大地提升了互联网的稳定性和扩展性。它允许不同自治系统之间进行高效的路由交换，并通过策略控制机制提供了灵活的管理手段。随着IPv6的普及和MPLS等技术的引入，BGP也不断演化以适应新的需求[20]。

2）DNS

DNS是互联网中用于将人类可读的域名转换为计算机可理解的IP地址的关键协议。从20世纪末到21世纪初，DNS

逐渐成熟并广泛应用于互联网中。DNS技术特征为：拥有分布式数据库，支持层次结构、多种记录类型，具备缓存机制，支持动态更新。20世纪90年代逐步成熟的DNS协议在互联网发展中扮演了至关重要的角色，其技术特征和里程碑事件不仅推动了域名解析技术的进步，也为后续的网络扩展和创新奠定坚实的基础。通过不断的改进和优化，DNS协议继续在全球互联网中发挥着关键作用[12]。

3.4 HTTP

21世纪初业界的核心技术是支持万维网（WWW）的超文本传输协议（HTTP）。HTTP是互联网中用于数据通信的基础协议。HTTP技术特征为：具有请求-响应模型、无状态性、头部信息、缓存控制，支持多种方法、持久连接等。21世纪初是HTTP协议快速发展和普及的时期。其技术特征和里程碑事件不仅推动了Web技术的进步，也为后续的网络扩展和创新奠定了坚实的基础。通过不断的改进和优化，HTTP协议继续在全球互联网中发挥着关键作用，支持各种复杂的Web应用和服务[21]。

3.5 HTTPS

21世纪00年代，业界核心的技术是HTTPS。HTTPS和TLS是互联网中用于保障数据传输安全性的关键技术。21世纪10年代是HTTPS/TLS协议快速发展和普及的时期，技术特征和里程碑事件显著推动了其在全球范围内的应用和扩展。HTTPS/TLS技术特征为：加密传输，需要身份验证，数据具有完整性，使用密钥交换，支持多版本、扩展机制等。值得一提的是：2013年的斯诺登事件、2015年由电子前沿基金会（EFF）和Mozilla等组织推动的HTTPS Everywhere运动，使HTTPS得到了广泛的部署。HTTPS/TLS协议在全球互联网中发挥着关键作用，保障数据传输的安全性和隐私性[21]。

4 互联网技术部署速率和创新

从互联网具体技术模块的角度看，可以观察到一个非常有趣的现象，有些创新型的技术虽然非常重要，但是其部署过程非常缓慢，例如：IPv6、DNSSEC、rPKI等；也有些黑马类的技术，其普及过程极其迅速，例如：NAT、HTTPS等。

4.1 部署速率比较分析

1）DNS安全扩展

DNSSEC是一项用于增强DNS安全性的技术，其主要目标是通过数字签名和加密来保护DNS查询和响应，防止DNS欺骗和缓存中毒等。DNSSEC规范于1997年发布，1998—1999年进行了改进和重新设计，2006年再次修订。2010年，根区域启用了DNSSEC签名，这标志着DNSSEC开始在全球范围内逐步部署和推广。但是，目前DNSSEC的部署和使用情况仍然相当有限。究其原因在于：DNSSEC复杂性大，存在兼容性问题，缺乏部署动力，并存在实施风险等[12]。

2）rPKI

rPKI的主要目标是提高互联网路由安全性，防止BGP劫持等攻击。2010年IETF定义了rPKI的基本框架和操作规范，2015年区域互联网注册管理机构（RIR）和互联网服务提供商（ISP）开始逐步部署rPKI，并推动其在全球范围内的采用。但直至今日，rPKI的部署和使用情况仍然相当有限。究其原因在于：rPKI复杂性大，存在兼容性问题，缺乏部署动力，实施存在风险等[13]。

3）NAT

NAT主要目的是解决IPv4地址短缺问题，提高网络安全性。1994年IETF定义了NAT的基本框架以及地址转换和端口映射操作规范，至此NAT就迅速得到普及。NAT快速普及的原因为：NAT简单易实现，安全性强，兼容性好，能够对部署者立即产生正面效果[22]。

4）HTTPS

HTTPS是一种在HTTP协议基础上加入SSL/TLS安全层的通信协议，其主要目的是提高数据传输的安全性，防止数据在网络传输过程中被窃取或篡改。1994年Netscape Communications公司开发了SSL协议。1995年Netscape发布了SSL 2.0和3.0版本，并开始在其浏览器中支持这些协议。1999年IETF标准化了SSL 3.0，并将其发展为TLS 1.0协议。2018年IETF发布了TLS 1.3版本，大幅度改进了安全性和性能，减少了连接延迟并移除了一些不安全的加密算法。HTTPS技术在全球范围内得到了快速部署和广泛应用，主要原因为：人们对于安全需求有所增加，主流浏览器支持HTTPS；搜索引擎优化促进HTTPS推广；HTTPS技术快速成熟。特别是免费证书颁发机构如Let's Encrypt的出现，使得获取和管理SSL/TLS证书更加便捷和低成本。此外，政策和法规的推动以及技术生态系统支持也是重要原因[21]。

综上所述，部署缓慢的技术有如下特点：

- 涉及范围广，需要从设备到软件再到基础设施的全面升级。
- 复杂性高，配置和管理复杂，需要人们具备专业知识和经验。
- 成本高，硬件和软件升级涉及高额成本。

• 现有系统依赖，许多应用和服务仍然基于旧协议设计，迁移需要时间和资源。

部署迅速的技术有如下特点：

• 涉及范围小，通常只需在局部进行配置或升级。

• 复杂性低，配置和管理相对简单，无须大规模基础设施升级。

• 成本低，硬件和软件升级成本较低，甚至不需要额外投资。

• 用户需求驱动，能够迅速满足用户或市场的需求，提供明显的短期效益。

4.2 IPv6过渡技术分析

IPv6技术也是典型的部署缓慢的技术，主要原因包括：需要对现有基础设施投资；无法和IPv4互联互通；IPv4仍有一个活跃的二级市场，供企业或用户满足需求；企业和用户缺乏强烈的动机去主动转向IPv6[22]。

1） 以双栈技术为主的传统IPv6过渡技术

传统的IPv6过渡技术的思路是“尽量使用双栈，必要时使用隧道”。然而，双栈方法虽然在技术上可行，但无法解决上述部署成本高、无法与IPv4互联互通等一系列重大问题。

2） 以翻译技术为主的新一代IPv6过渡技术

新一代的IPv6过渡技术（如RFC6145、RFC6052、RFC6146、RFC7599等）提出了“尽量使用翻译，必要时使用双重翻译或隧道，外特性为双栈”的思路。通过翻译技术，IPv6单栈服务器和IPv6单栈客户机可以与现有的IPv4互联网互联互通。这样一来，部署翻译器的网络可以率先过渡到IPv6单栈，同时保证与IPv4互联互通。翻译技术的优点为：

• 涉及范围小，通常只需在局部进行配置或升级。

• 复杂性低，配置和管理相对简单，无须大规模基础设施升级。

• 成本低，硬件和软件升级成本较低，甚至不需要额外投资。

• 用户需求驱动，能够迅速满足用户或市场的需求，提供明显的短期效益。

• 快速部署，通过翻译器，IPv6单栈设备可以迅速与现有的IPv4网络互联互通，加快了IPv6的普及速度。

• 安全性强，由于“木桶效应”，双栈系统的总体安全性是IPv4或IPv6中安全性差的那一个，IPv6单栈系统可以充分利用IPv6增强的安全性。

5 新的挑战

虽然当今互联网基本上保持了TCP/IP的“分布式”“无连接”“尽力而为”“端对端”和“开放性”的技术特征，但“去分布式”“去无连接”“去尽力而为”和“去端对端”的异化也带来了新的挑战[14]。

5.1 网络韧性

在当今全球化和数字化的世界中，网络基础设施的韧性至关重要。网络韧性指的是系统在面对干扰、故障或攻击时仍能保持正常运行的能力。然而，地缘政治和自然灾害对网络韧性提出了严峻的挑战。

地缘政治紧张局势可能导致网络攻击、信息封锁和互联网服务中断。例如，国家之间的网络战争可能通过DDoS、恶意软件等手段破坏对方的关键基础设施。不同国家和地区的互联网管理政策和法规差异，可能导致跨境数据传输受阻，影响全球网络的连通性和韧性。某些国家可能实施严格的网络封锁或内容过滤政策，削弱了网络的整体韧性。地缘政治因素也可能影响关键资源的分配和使用，如海底光纤电缆的铺设和维护。这些基础设施的中断或破坏将直接影响网络的韧性。

台风、飓风、暴雨等自然灾害可能导致电力中断、设备损坏和通信线路故障，影响网络的正常运行。地震、海啸等地质活动可能破坏关键的通信基础设施，导致大规模的网络中断。长期的气候变化可能增加极端天气事件的频率和强度，进一步考验网络基础设施的韧性。

为了应对上述挑战，恢复和加强分布式网络结构是关键。分布式网络结构具有更高的冗余性、灵活性和弹性，能够在局部故障或攻击发生时保持整体网络的正常运行。一般来讲，韧性网络具有以下特征：

1）多重路径

分布式网络通过多条路径传输数据，确保即使某些路径受阻或中断，数据仍能通过其他路径传递到目的地。这种设计可以有效应对自然灾害和地缘政治因素导致的局部网络故障。

2）去中心化

分布式网络没有单一的控制中心，数据存储和处理分散在多个节点上。这种结构减少了单点故障的风险，提高了系统的整体韧性。

3）动态调整

分布式网络能够根据实时情况动态调整数据传输路径和负载分配，确保在面对突发事件时能够迅速响应并恢复正常运行。

4）本地化服务

通过增加本地化的数据中心和服务节点，减少对远程资源的依赖，提高网络在局部灾害或冲突发生时的自愈能力。

5） 跨国合作

国际社会需要加强合作，共同制定和实施全球性的网络安全和韧性标准，确保关键基础设施的互联互通和稳定运行。

5.2 人工智能

近年来，以ChatGPT为代表的生成式人工智能对互联网和TCP/IP体系结构提出了新的挑战[23-24]。

1）高性能网络的挑战

AI模型通常需要处理大量数据，尤其是在训练阶段。高带宽可以确保数据快速传输，从而提升训练效率。低延时对于实时应用（如对话式AI）至关重要；高延时会影响用户体验和系统响应速度；网络抖动会导致数据传输的不稳定，影响模型训练和推理的一致性；高丢包率会导致数据传输的不完整，影响模型的准确性和性能。以英伟达为代表的图形处理器（GPU）集群为例，NVLink被用于Nvidia GPU之间的通信，因为它提供了远超传统高速串行计算机扩展总线标准（PCIe）接口的带宽和更低的延时，对于需要大规模并行计算的AI任务（如深度学习训练）至关重要。Fiber Channel主要用于数据中心内部的存储网络，因为它提供了极低延时和高带宽，确保数据传输的高效性和稳定性，对于需要快速访问大量数据的AI应用尤为重要。TCP/IP被广泛用于数据中心之间的通信，主要原因在于其标准化和广泛支持。它可以在不同硬件和软件平台上运行，适合大规模分布式系统。目前的技术在近期无法使TCP/IP一统天下。

2）分裂化的挑战

随着国际关系的复杂化，以及各国对数据主权、隐私保护以及网络安全的重视程度的提高，互联网的分裂问题日益严重。具体到人工智能的服务领域，许多公司开始根据源IP地址决定是否提供服务，这种做法在未来可能会产生深远的影响。例如，从2024年7月9日起，OpenAI严格限制其API服务的IP地址范围。这意味着只有特定国家或地区的用户才能访问和使用OpenAI提供的人工智能服务。从长远看，这可能会导致这些地区在技术创新方面有所落后，全球范围内的研究和开发合作将变得更加困难，最终影响技术进步的速度和质量。

3）集中化的挑战

人工智能，特别是大型基础模型（如语言模型、图像识别模型等）的训练，需要大量的计算资源（如GPU）和能源。这些资源的成本非常高昂，只有少数超大公司才能负担得起，从而导致了人工智能领域的中心化趋势。其长远影响在于：大公司对于技术的垄断阻碍创新；超大公司控制了大量用户数据，可能导致数据隐私问题加剧；中心化的系统更容易成为黑客攻击的目标，一旦被攻破，后果将非常严重等；中心化趋势可能进一步加大数字鸿沟，使得资源和技术更集中在发达国家和地区。

4）可信性的挑战

大语言模型在处理自然语言任务时，有时会生成看似合理但实际上并不准确或可信的回答。这种现象被称为“幻觉”。与此同时，互联网用户还面临来自黑客攻击和劫持的威胁，这使得区分大语言模型生成的幻觉和恶意行为变得更加复杂。其长远影响在于：如果用户无法区分大语言模型生成的幻觉和恶意行为，则可能会对人工智能服务失去信任，影响其广泛应用；企业提供的人工智能服务如果频繁出现幻觉或被黑客攻击，可能会导致品牌声誉受损；黑客攻击和劫持可能导致用户数据泄露，造成严重的隐私问题，大语言模型生成的幻觉可能误导用户做出错误决策，甚至引发安全事件。因此，需要通过网络对联网实体（不管是真实的人类，物联网设备和“机器人”）通过真实的网络地址建立更强的认证和信任机制。

6 结束语

在本文中，我们回顾了TCP/IP的历史演进、TCP/IP的核心思想以及各技术模块的发展过程，通过深入分析，我们得出以下结论：

在当前快速变化的技术环境中，TCP/IP作为互联网的基石，其核心思想依然具有重要意义。未来的发展应当坚持“守正创新”的原则：一方面，保持TCP/IP“分布式”“无连接”“尽力而为”“端对端”和“开放性”的基本技术特征；另一方面，大胆创新，适应人工智能等新技术带来的需求。通过这种方式，我们可以在确保网络基础设施稳定性的同时，推动技术进步和创新。

随着IPv4地址资源的枯竭和互联网应用需求的不断增长，IPv6作为下一代互联网协议，为技术创新提供了坚实的基础。IPv6不仅解决了地址空间不足的问题，还能够为移动互联网、物联网和人工智能提供不受限制的创新空间。

参考文献

[1] “863”计划通信技术主题总体技术研究组. BIP-ISDN概念研究报告[R]. 1995

[2] LEINER B, CERF V, CLARK ,et al. A brief history of the Internet

[EB/OL]. (2022-02-22) [2024-10-13]. https://www.internethalloffame.org/brief-history-internet
[3] 李星, 包丛笑. 五十年互联网技术创新发展的回顾与思考 [J]. 汕头大学学报(人文社会科学版), 2019, 35 (12): 5-12
[4] RUSSELL A L. OSI: the Internet that wasn't [EB/OL]. (2013-07-29) [2024-10-04]. https://spectrum. ieee. org/osi-the-internet-that-wasnt
[5] CLARK D. Designing an Internet- information policy [M]. Cambridge: The MIT Press, 2018
[6] U. 布莱克. 计算机网络 - 协议、标准与接口 [M]. 北京: 人民邮电出版社, 1990
[7] MigiTing. ISDN与Internet [M]. 北京: 机械工业出版社, 1997
[8] 邢秦中. ATM通信网 [M]. 北京:人民邮电出版社, 1998
[9] SCHHATT, STAN. Linking LANs - a micro manager's guide [M]. PA: TAB BOOKS, 1991
[10] 和永明, 陈地虎. OSI协议和计算机网 [M]. 北京:电子工业出版社, 1994
[11] HALEPLIDIS E, PENTIKOUSIS K, DENAZIS S, et al. Software-defined networking (SDN): layers and architecture terminology [S]: RFC7426. 2015
[12] IETF. Domain name system [EB/OL]. [2024-10-10]. https://www.ietf.org/technologies/dns/
[13] MANDERSON T, VEGODA L, KENT S. Resource public key infrastructure (RPKI) objects Issued by IANA RFC6491 [EB/OL]. [2024-10-06]. https://datatracker.ietf.org/doc/rfc6491/
[14] 晃通, 宫永直树, 岩田淳. 图解OpenFlow [M]. 北京: 人民邮电出版社, 2016
[15] BUSI L, ALLAN D. Operations, administration, and maintenance framework for MPLS-based transport networks RFC6371 [EB/OL].[2024-10-03]. https://datatracker.ietf.org/doc/rfc6371/
[16] FILSFILS C, LEDDY J, VOYER D, et al. Segment routing over IPv6 (SRv6) network programming RFC8986 [EB/OL].[2024-10-03]. https://datatracker.ietf.org/doc/rfc8986/
[17] BERNET Y, FORD P, YAVATKAR Y, et al. A framework for integrated services operation over diffserv networks RFC2998 [EB/OL].[2024-10-03]. https://datatracker.ietf.org/doc/rfc6371/
[18] CARPENTER B. Internet Transparency RFC2775 [EB/OL]. [2024-10-03]. https://datatracker.ietf.org/doc/rfc2775/
[19] IETF. Internet standards process [EB/OL]. [2024-10-03]. https://www.ietf.org/process/process/
[20] REKHTER Y, LI T. A border gateway protocol 4 (BGP-4) RFC1771 [EB/OL]. [2024-10-05]. https://datatracker.ietf.org/doc/rfc1771/
[21] FIELDING R, NOTTINGHAM M, RESCHKE J. HTTP semantics RFC9110 [EB/OL]. [2024-10-05]. https://datatracker.ietf.org/doc/rfc9110/
[22] 李星, 包丛笑. 新一代IPv6过渡技术——IPv6单栈和IPv4即服务 [M]. 北京: 科学出版社, 2024
[23] WILLIAM J, DARK VINTON G, KLEINWACHTER C W. Internet fragmentation: an overview [EB/OL]. [2024-10-10]. https://www3. weforum. org/docs/WEF_FII_Internet_Fragmentation_An_Overview_2016.pdf 2016
[24] ISOC. Artificial intelligence – Internet society [EB/OL]. [2024-10-10]. https://www.internetsociety.org/issues/past-categories/ai/

作 者 简 介

李星，清华大学教授、CERNET网络中心副主任；主要研究领域为计算机网络体系结构、通信技术等；作为负责人完成多项国家级项目，获中国科技进步一等奖、中国科技发明二等奖、通信学会科学技术奖一等奖等，入选国际互联网名人堂，获互联网波斯塔尔奖；作为联合作者撰写11个IETF RFC，发表学术论文300余篇，获国家发明专利40余项。

包丛笑，清华大学副教授；主要研究领域为计算机网络体系结构、IPv6过渡技术和网络测量；作为技术骨干完成多项国家级项目，获中国通信学会科学技术奖一等奖；作为联合作者撰写9个IETF RFC，发表学术论文30余篇，获国家发明专利40余项。

数据中心液冷散热技术及应用

Technology and Application of Liquid Cooling Heat Dissipation in Data Centers

严劲/YAN Jin[1]，景焕强/JING Huanqiang[2]，
张子骜/ZHANG Ziao[1]，刘帆/LIU Fan[2]
（1. 中国电信智能云网调度运营中心，中国 北京 100010；
2. 中兴通讯股份有限公司，中国 深圳 518057）
(1. Intelligent Cloud Network Operating Center, China Telecom Group, Beijing 100010, China;
2. ZTE Corporation, Shenzhen 518057, China)

DOI：10.12142/ZTETJ.202406013
网络出版地址：http://kns.cnki.net/kcms/detail/34.1228.TN.20241212.1054.004.html
网络出版日期：2024-12-12
收稿日期：2024-10-20

摘要：数字技术的创新演进与蓬勃发展，推动算力需求持续提升，数据中心能耗呈指数型增长。在可持续发展、“双碳”、新型数据中心等政策理念指引下，数据中心制冷技术正式迈入液冷阶段。首先从芯片、设备、机柜散热诉求，机房节能诉求等多个维度，深入探讨液冷技术的必要性与优势，同时针对多种液冷技术方案从架构、原理、关键组成等方面进行深入分析。其次，通过散热能力、节能效果、维护性、技术成熟度等方面的综合对比，短中期单相冷板式液冷将更具优势。最后，探讨了当前数据中心液冷在可靠性、散热强化、低成本等维度的研究趋势。

关键词：散热技术；机房节能；液冷散热；冷板式液冷；浸没式液冷

Abstract: The innovative evolution and vigorous development of digital technology have led to a continuous increase in computing power demand and an exponential growth in data center energy consumption. Under the guidance of policies such as sustainable development, "dual carbon" goals, and new-style data centers, data center refrigeration technology has officially entered the liquid cooling stage. This paper first discusses the necessity and advantages of liquid cooling technology from the aspects of chip/equipment/cabinet heat dissipation requirements and energy saving requirements, and based on the architecture, principles, and key components, a variety of liquid cooling technology solutions are deeply analyzed. Secondly, through a comprehensive comparison of heat dissipation capacity, energy-saving effect, maintainability, and technical maturity, it can be seen that single-phase cold plate liquid cooling will have more advantages in the short and medium term. Finally, the current research trends of data center liquid cooling in terms of reliability, heat dissipation enhancement, and low cost are discussed.

Keywords: heat dissipation technology; computer room energy saving; liquid cooling heat dissipation; cold plate liquid cooling; immersion liquid cooling

引用格式：严劲, 景焕强, 张子骜, 等. 数据中心液冷散热技术及应用 [J]. 中兴通讯技术, 2024, 30(6): 84-91. DOI: 10.12142/ZTETJ.202406013
Citation: YAN J, JING H Q, ZHANG Z A, et al. Liquid cooling heat dissipation technology and application in data centers [J]. ZTE technology journal, 2024, 30(6): 84-91. DOI: 10.12142/ZTETJ.202406013

1 应用背景

1.1 节能政策驱动

随着数字技术的创新演进，云计算、大数据、人工智能（AI）、元宇宙等信息技术和实体经济深度融合，推动数字经济持续快速增长。数据中心是数字经济基础设施的底座。数据量爆发式增长带动数据中心市场快速增长。数据显示，截至2023年底，中国在用数据中心机架总规模达到810万标准机架。作为“能耗大户”，数据中心的耗电量不断刷新纪录，数据中心的总用电量约占全社会用电量3%。在可持续发展、“碳达峰、碳中和”、新型数据中心等政策理念指引下，国家及地方政府相继出台相关政策，对数据中心电源使用效率（PUE）提出更高要求。

工业和信息化部于2021年7月印发《新型数据中心发展三年行动计划（2021—2023年）》，明确到2023年底，新建大型及以上数据中心PUE降低到1.3以下，东数西算枢纽节点及寒冷地区力争降低到1.25以下[1]。国家发展改革委在2021年11月印发《贯彻落实碳达峰碳中和目标要求 推动数据中心和5G等新型基础设施绿色高质量发展实施方案》，进一步明确“到2025年，新建大型、超大型数据中心PUE降到1.3以下，国家枢纽节点降至1.25以下”[2]。东数西算工程八大枢纽节点，要求东部地区PUE目标不超过1.25，西部地区不超过1.2，能效指标更加严格。

在典型数据中心能耗占比中，制冷系统占比达到24%以上，是数据中心辅助能源中占比最高的部分。因此，降低数据中心PUE的关键在于采用更加高效节能的制冷方案。

近年来，为了降低制冷系统电能消耗，业内对机房制冷技术进行了持续的创新和探索，如间接蒸发冷却、冷板式液冷、浸没式液冷[3]等。其中，间接蒸发技术的PUE可达1.25，液冷技术则利用液体的高导热、高传热特性，在进一步缩短传热路径的同时充分利用自然冷源，可以实现数据中心PUE低至1.1的极佳节能效果。得益于绿色节能优势，近年来液冷技术也成为国家及地方政策明确鼓励采用的重要节能技术，如表1所示。

1.2 高散热诉求

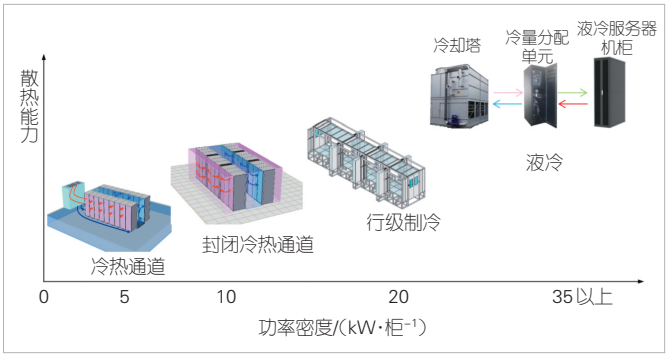
算力的持续增加促进通信设备性能不断提升，市场主流芯片功耗和热流密度也在持续攀升，中央处理器（CPU）散热设计功耗已达350~500 W。AI技术的快速发展推动图形处理器（GPU）需求增长，GPU散热设计功耗已超过800 W。芯片功率密度的持续提升直接制约着芯片散热和可靠性。

芯片功率密度的攀升同时带来整柜功率密度持续增长。8 kW以上单机柜功率密度成为目前新建数据中心的主流选择。为提升市场竞争力，人们通过升级改造的方式来提高单柜功率密度。目前，通算最大功率密度已超过30 kW/柜，如图1所示。智算功率上升更快，已达100 kW/柜。整机柜功率密度的提升对机房制冷技术提出了更高的要求。传统风冷系统受数据中心建筑面积与单位运营成本等因素的影响散热上限一般为20 kW/柜[4]，越来越难以为继。液冷技术采用液体替代空气作为冷却介质，将液体直接或间接接触发热器件，可使散热效率大幅提升，能够有效满足单点、整机柜、机房的高散热需求。

2 液冷技术分类

根据热器件是否与冷却液接触，液冷技术可以分为直接接触式和间接接触式两种：直接接触式是指将冷却液体与发热器件直接接触散热，这类液体包括单相浸没式液冷、两相浸没式液冷、喷淋式液冷；间接接触式是指冷却液体不与发热器件直接接触，通过散热器间接散热，这类液体包括单相冷板式液冷、两相冷板式液冷。

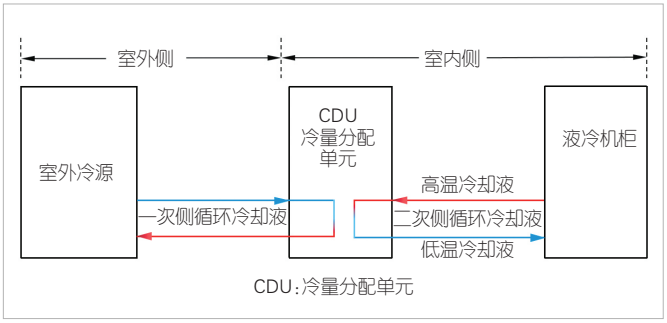
液冷系统通用架构如图2所示。其中，室外侧包含室外冷源、一次侧冷却液，室内侧包含冷量分配单元（CDU）、二次侧冷却液以及液冷机柜。该液冷系统的基本原理是：二次侧冷却液在机柜内吸收设备热量，并通过CDU内的换热器将热量传递给一次侧冷却液，一次侧冷却液通过室外冷源



▲图1 机柜功率密度与制冷方式

▼表1 液冷数据中心政策

发布时间	发布主体	政策文件	液冷政策
2021-07	工业和信息化部	《新型数据中心发展三年行动计划（2021—2023年）》	鼓励应用液冷等高效制冷系统
2021-12	国家发展改革委、工业和信息化部、国家能源局	《贯彻落实碳达峰碳中和目标要求 推动数据中心和5G等新型基础设施绿色高质量发展实施方案》	支持数据中心采用新型机房精密空调、液冷、机柜式模块化等方式
2022-07	北京市经济和信息化局	《北京市推动软件和信息服务业高质量发展的若干政策措施》	对数据中心转型为算力中心或涉及液冷应用的，按照固定资产投资的30%进行奖励
2022-12	重庆市通信管理局、重庆市经济和信息化委员会等	《重庆市信息通信行业绿色低碳发展行动实施方案（2022—2025年）》	积极应用液冷型IT设备，提高数据中心IT设备能效
2022-12	成都市经济和信息化局	《全国一体化算力网络成渝国家枢纽节点（成都）推进方案》	鼓励数据中心液冷利用等先进供冷技术
2023-03	财政部 环境部 工信部	《绿色数据中心政府采购需求标准（试行）》	数据中心相关设备和服务应当优先选用新能源、液冷等高效方案
2023-03	广东省发展和改革委员会、广东省能源局等	《广东省绿色高效制冷行动计划（2023—2025）》	鼓励使用液冷服务器、自动喷淋等高效制冷系统，因地制宜采用自然冷源等制冷方式，大幅提升数据中心能效水平



▲图2 液冷系统通用架构图

最终将热量释放到大气环境中，完成散热。

1）室外冷源：可选择开式/闭式冷却塔、干式冷却器等，冷源的选择应根据所在地的场地、气象、水电等因素综合考虑。

2）一次侧冷却液：常用的液冷液有去离子水、乙二醇水溶液、丙二醇水溶液等，并配合具有一定缓蚀、杀菌、阻垢功能的化学药剂使用。冷却液的选择需要根据液体热物性、部署地理位置及气候条件等综合考虑。

3）CDU：按布置形式可分为集中式与分布式。其中，集中式CDU布置在机柜外，为多台液冷机柜提供冷量，易于集中化部署和管理；分布式CDU布置在液冷机柜内部，每台机柜对应一个CDU，易于机柜功耗匹配。

二次侧冷却液、液冷机柜及内部液冷设备在不同液冷技术形态中略有差异，在后续章节中我们会具体介绍。

2.1 单相冷板式液冷

单相冷板式液冷通过液冷板将发热器件的热量间接传递给液冷板中的二次侧冷却液。二次冷却液在设备吸热和CDU放热过程不发生相变。根据液冷板覆盖范围，这种液冷可以分为局部液冷或全液冷：局部液冷通常仅覆盖高功耗器件，一般带走设备70%左右的热量，剩余30%热量仍需通过机房空调或液冷背门以风冷的形式带走；全液冷需要根据通信设备硬件架构和结构布局定制化设计液冷板，以覆盖所有发热器件。单相冷板式液冷系统架构如图3所示，液冷机柜内包含分液器、液冷板、流体连接器、液冷管路、漏液检测传感器等。

1）二次侧冷却液：二次侧热量载体以去离子水、乙二醇水溶液、丙二醇水溶液为主，根据具体场景进行选择。二次侧冷却液需要定期检测PH、浊度、残留物、细菌等参数，并符合相关标准要求。

2）单相冷板CDU：可分为集中式和分布式。其中，集中式CDU布置在机柜外，每列机柜布置一台或几台CDU，实现主用和备份关系，需要部署二次侧管网，并考虑各液冷机柜间的流量分配；分布式CDU安装在液冷机柜内，免二次侧管路部署，可根据机柜功耗灵活部署。

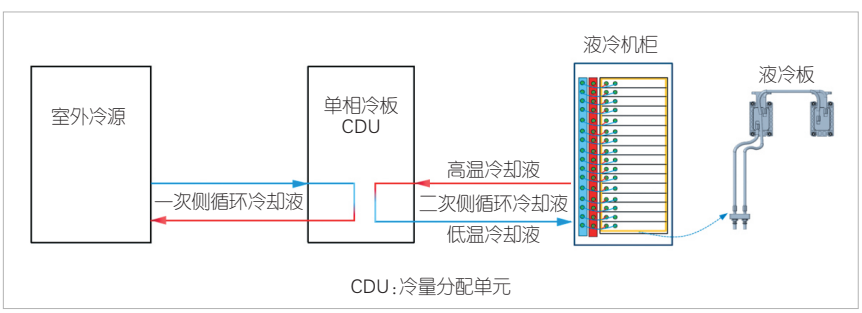
3）分液器：用于机柜内流量分配与收集，将低温二次侧冷却液分配到各设备节点，并收集与液冷板换热升温后的冷却液。其设计选型过程中需要保证流量分配需要的均匀性，并结合机柜空间、重量等要求综合考虑分液器的体积。

4）液冷板：液冷板设计需要根据设备芯片功耗进行芯片冷板设计、根据芯片布局及单板结构空间设计冷板连接管路路由，具有一定的定制化特性。但在进行设计时应尽量保证内部零件的通用性，如内部翅片规格、进出口规格应尽可能一致，以降低成本。此外，液冷板的设计还需要综合考虑实际功耗、工作压力、流速等。

5）流体连接器：可实现无泄漏通断，在设计选型时需要综合考虑工作流量、温度、压力、流阻特性、安装方式、直插/盲插、接口规格等。

6）液冷管路：二次侧冷却液流通通路，参与液冷机柜内各设备节点的流量–流阻分配；液冷管路设计选型需要考虑材料兼容性、流速、管路布置、安装方式、流量分配设计等。

7）漏液检测传感器：针对沿液冷板、液冷管路、分液器等可能出现液体泄漏的位置或路径布置，及时检测泄漏状态，并触发漏液告警策略，及时告知运维人员发现漏液事故，便于及时处理，有效地保护液冷系统与机房安全。漏液检测传感器可分为检测线、检测带、光电式、电极式、浮子式等，适用于不同的泄漏位置和泄漏场景。



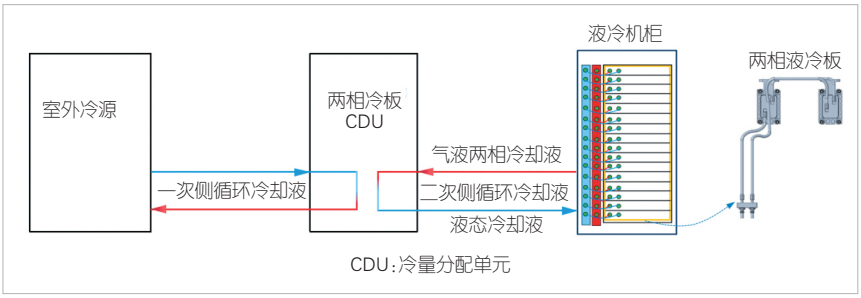
▲图3 单相冷板式液冷系统架构

单相冷板式液冷技术对通信设备和机房基础设施改动较小，业内已具备多年研究积累，目前技术成熟度最高，它已成为满足芯片高热流密度散热需求、提升数据中心能效、降低总体拥有成本（TCO）的有效方案。

2.2 两相冷板式液冷

两相冷板液冷系统架构与单相液冷板

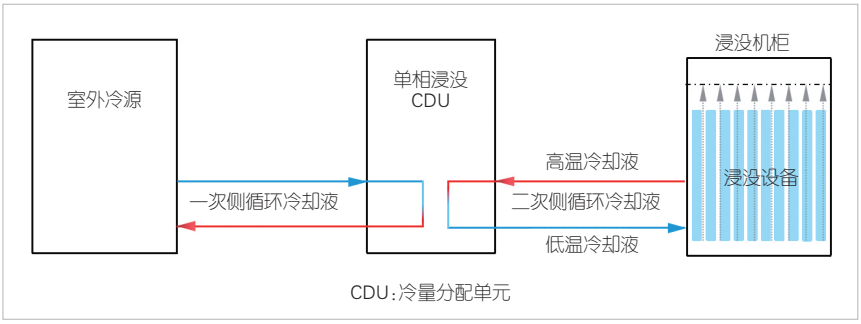
液冷相似，其系统架构如图4所示。所不同的是二次侧冷却液在设备内通过液冷板吸热发生汽化，在CDU内冷凝为液态，充分利用了冷却液的相变潜热，综合散热能力更强，可达300 W/cm^{2}以上。由于运行过程中系统内冷却液发生相变，两相冷板液冷系统的压力会高于单相冷板液冷，其二次侧冷却液、液冷板、流体连接器、液冷管路等为了适配系统压力也要满足一定的特殊化要求。



▲图4　两相冷板式液冷系统架构

1）二次侧冷却液：以制冷剂、氟化液等低沸点工质为主，在选型时主要考虑热物性、环保性、安全性、工作温区和压力、材料兼容性等因素。

2）两相冷板CDU：两相冷板液冷系统压力等级通常较高，其压力控制系统区别于单相系统，一般采用温控型压力控制方案。同时，两相CDU补液系统在设计时也需要考虑工质充注量对于系统压力的影响。



▲图5　单相浸没式液冷系统架构(卧式)

3）两相液冷板：其结构与单相液冷板相似，在设计时需要重点考虑冷板承压能力，增加汽化核心、促进气泡脱离以提升散热性能，常见的方案有表面微处理、多孔介质填充等。

4）两相流体连接器：高压系统对流体连接器的插拔操作和带压维护都提出了很高的要求。目前螺纹旋拧连接器能够较好地满足需求。

5）液冷管路：考虑系统压力及气相工质泄漏风险，优选金属软管或汽车空调橡胶管。

两相冷板式液冷核心技术的优势在于能够满足超高热流密度散热需求，但现阶段技术成熟度仍较低，相关产业链还有待完善。

2.3 单相浸没式液冷

单相浸没式液冷通过将发热元件浸没在冷却液中，直接吸收设备产生的热量。卧式浸没液冷系统架构如图5所示，通信设备竖插在浸没机柜内，二次侧低温冷却液由浸没机柜底部流入。二次侧冷却液在循环散热过程中始终维持液相。

1）二次侧冷却液：单相浸没技术通常使用高沸点的冷却液。这类冷却液不发生相变，同时需要具有高绝缘、低黏度以及良好的兼容特性，例如氟碳化合物和碳氢化合物（矿物油、合成油等）。

2）浸没机柜：现阶段应用较多的为卧式机柜（通常称为TANK），业内常用的尺寸规格覆盖12U~54U。为了实现卧式架构下的流量均衡性，TANK底部需配置均流板。冷却液由底部进入，经均流板分液后流入设备。为便于通信设备的安装和维护，TANK设计需要有一定的槽位导向和固定功能。同时，TANK上盖与腔体之间需要具备良好的密封性，防止运行过程中冷却液耗散。

3）单相浸没CDU：单相浸没液冷系统在维护过程中需要打开TANK上盖，系统直接与机房环境连通，属于一种“半开式”系统，因此其CDU设计对循环泵、系统过滤、冷却液监控等要求更高。

单相浸没液冷实现了100%液体冷却，无须配置风扇，可使机房极致节能、静音。单相浸没液冷在应用时需要将通信设备完全浸没在冷却液中，所有材料、器件均需要重新选型评估，并开展兼容性测试验证以保证应用的可靠性。由于不导电液体热物性普遍较差且液体流速低，因此单相浸没液冷散热能力普遍较低，这在一定程度上制约了其推广应用。

根据浸没机柜形态，单相浸没式液冷可以进一步细分为卧式浸没和立式浸没。传统卧式浸没液冷设备维护时需要打开TANK上盖，并配备可移动机械吊臂或专业维护车以实现设备的竖直插拔，维护复杂度高、耗时长，且开盖维护过程存在一定的冷却液挥发问题，增加了运行成本。为了解决这一问题，业内将浸没机柜形态调整为立式架构，即单相立式浸没液冷，如图6所示。立式浸没机柜架构与冷板式相似，但通信设备本身需要实现板级密封功能，兼具冷板式液冷的维护便利性和浸没式液冷的节能优势。

2.4 两相浸没式液冷

两相浸没液冷二次侧冷却液在设备内吸热由液态转化为气态，通过冷凝器冷凝放热由气态转化为液态。这种液冷技术充分利用液体的相变潜热，散热能力相比于单相浸没显著提升。需要指出的是，两相浸没液冷同样存在卧式和立式两种技术形态。

两相卧式浸没二次侧冷却液仅在浸没腔体内部循环。浸没腔体的顶部为气态区，底部为液态区。冷却液吸收设备热量后发生相变，即液态冷却液变为气态冷却液。气态冷却液汇聚到浸没腔体顶部，与安装在顶部的冷凝器发生换热后冷凝为低温液态冷却液，随后在重力作用下回流至腔体底部，实现对通信设备的散热，如图7所示。

两相立式浸没将每个设备节点作为一个独立的小型浸没腔体，可有效避免相变冷却液的运维耗散问题，且架构兼容性更优、维护操作更便捷。因此，现阶段两相浸没以立式架构为主要研究方向。两相浸没立式系统架构如图8所示，它包含二次侧冷却液、密封壳体、两相沸腾散热器等关键部件。

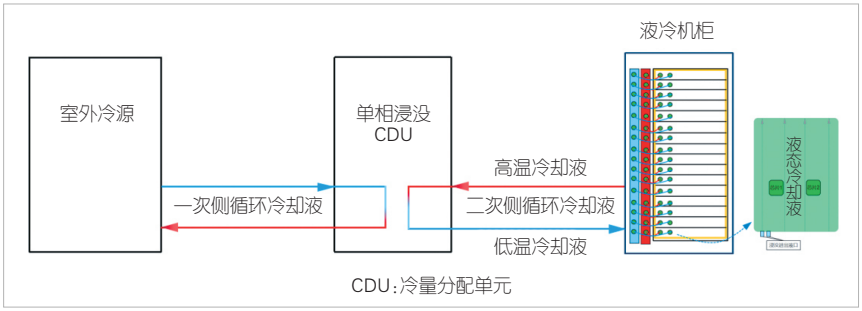
1）二次侧冷却液：考虑密封壳体的承压设计，目前主要选用低沸点氟碳类工质。二次侧冷却液需要根据具体场景进行选择，并主要考虑热性能、环保安全性能、工作温区和压力、材料兼容性等因素。

2）密封壳体：通信设备节点全密封设计，节点内部充满冷却液。工作时壳体上部为气体，下部为液体，通过流体连接器与CDU形成气液循环。密封壳体的关键点在于设备电、网、液接口处的密封设计。

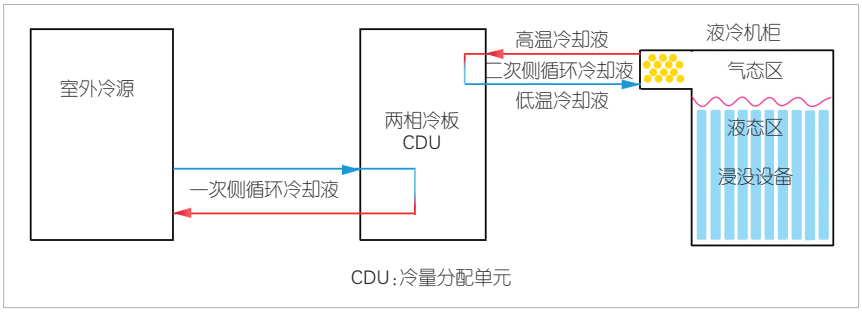
3）两相沸腾散热器：通过界面材料与芯片接触，将芯片产生的热量通过冷却液的相变带走。这类散热器一般采用多孔介质设计方案，以增加汽化核心和散热面积[5]。

两相浸没液冷兼具高节能、高散热的技术优势，可同时满足高功率芯片的散热需求，实现机房极致节能效果。但现阶段该技术仍在试点研究中，其密封可靠性、系统控制稳定性等有待持续优化。

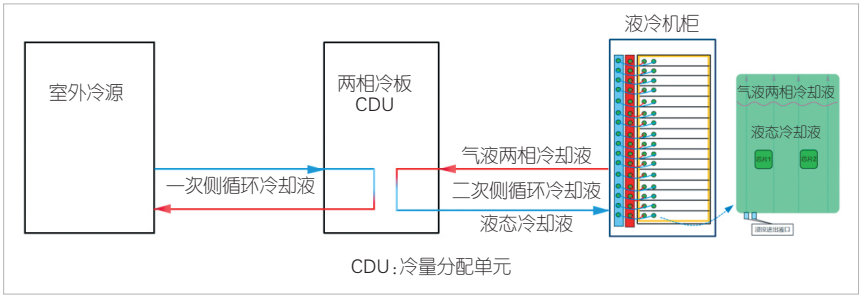
两相浸没液冷兼具高节能、高散热的技术优势，可同时满足高功率芯片的散热需求，实现机房极致节能效果。但现阶段该技术仍在试点研究中，其密封可靠性、系统控制稳定性等有待持续优化。



▲图6　单相浸没式液冷系统架构（立式）



▲图7　两相浸没式液冷系统架构（卧式）



▲图8　两相浸没式液冷系统架构（立式）

2.5 喷淋式液冷

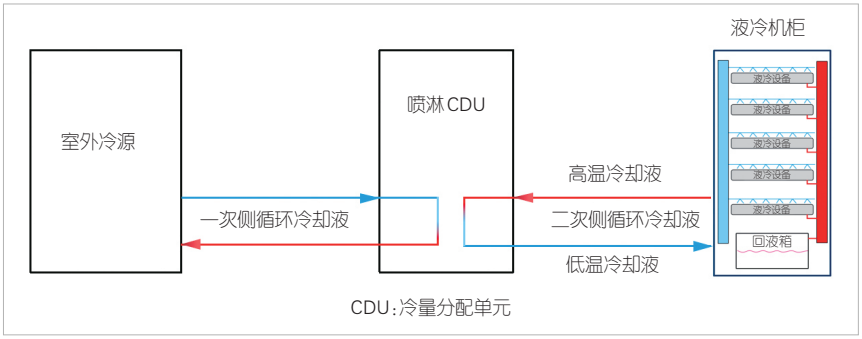
喷淋式液冷属于直接接触式液冷。二次侧冷却液由顶部进入服务器，在重力或系统压力的作用下，通过喷淋板精准喷淋发热器件，冷却液直接与发热器件接触，通过对流换热为器件散热，如图9所示。为了实现精准喷淋与有效散热，液冷机柜及设备需要一定的特殊化设计。

1）二次侧冷却液：通常为不导电液体，可以是油基或氟碳类，换热过程不发生相变。

2）液冷设备：上盖集成喷淋腔体和喷淋孔，可根据器件功耗、布局、尺寸设置不同的喷淋孔大小、位置、密集程度等。

3）液冷机柜：设备内喷淋会有一定的冷却液飘逸，为了避免冷却液损耗，以及机房环境污染，液冷机柜需要具备一定的密封性。

4）储液箱：一般放置于喷淋机柜的底部，利用重力收



▲图9　喷淋式液冷系统架构

集吸热温升后的冷却液，当系统出现异常情况时（如发生泄漏），也可收集泄漏液体，增加系统运行的稳定性和可靠性。

喷淋式液冷实现了100%液冷，使PUE优于单相冷板液冷。通过喷淋结构，这种液冷技术可实现对高功率芯片的精准喷淋，使流经芯片的液体流速得到提升。这种液冷技术的散热能力略高于传统单相浸没液冷。因此，喷淋液冷可以看作是实现冷板式液冷节能、单相浸没液冷散热的折中方案。

2.6 液冷技术综合对比

算力攀升驱动数据中心液冷市场需求保持逐年增长的态势。业内多条液冷技术路线快速发展，针对不同应用场景各具优势，如表2所示。其中，单相冷板式液冷在液冷数据中心的应用占比达90%以上，是现阶段及未来一段时间业内主流的液冷技术方案。单相浸没式液冷节能优势更突出，且近年来该技术逐步趋于成熟，相关产业链快速发展完善，小规模商用不断推进。此外，喷淋式、两相冷板式、两相浸没式这3种液冷方案的技术研究和产业生态尚需完善。

3 液冷技术展望

数据中心液冷正处于快速发展阶段。随着液冷技术的规模化应用，各类问题也逐渐暴露出来。例如：冷板式液冷水基工质泄漏导致设备短路烧毁；单相浸没式液冷散热能力受液体流速约束，散热能力表现较弱，无法满足更高功耗CPU/GPU的散热需求[6]；液冷系统制冷量未随负载变化及时调控，导致节能收益不明显；现阶段液冷数据中心的建设成本高等。这些均在一定程度上制约了液冷技术在数据中心领域的应用。为了解决这些问题，业界一直在持续探索研究，以提升数据中心液冷技术在安全可靠、散热能力、建设成本等方面的优势。

▼表2　不同液冷技术方案对比

对比项	单相冷板式	两相冷板式	单相浸没式	两相浸没式	喷淋式
初投资	5	3	3	2	3
运营成本	2	2	4	5	3
节能效果	2	2	4	5	3
散热能力	4	5	2	4	2
噪音程度	3	3	5	5	4
环境影响	5	5	3	2	3
维护性	5	4	2	3	2
空间利用率	4	4	2	5	3
技术成熟度	5	2	3	2	3

注1:得分5表示最优；
注2:单相浸没式以卧式架构为对比技术方案；
注3:两相浸没式以立式架构为对比技术方案。

3.1 非水冷板式液冷

单相冷板式液冷一般采用水基工质作为二次侧冷却液，但水基工质存在腐蚀、泄漏导电等应用可靠性风险。除了基础的机械结构防泄漏外，中兴通讯创新性地提出非水冷板式液冷技术，将二次侧冷却液由水基工质更换为氟碳类或油基不导电液体，从冷却液本身解决泄漏导电问题。非水冷板式液冷架构与单相冷板式液冷相同。

非水冷板式液冷方案配合机械防泄漏结构设计，能够实现对液冷系统的多维度泄漏防护，真正做到泄漏有效防护、不损伤设备，且保留了冷板式液冷的高散热优势，能够满足现阶段各类通信设备的散热需求。同时，由于氟碳类、油基工质均属于大分子化合物，很难被微生物所分解，因此，非水系统中微生物腐蚀导致的风险会大大降低。

非水冷板式液冷因工质更换，其系统方案在设计过程也需要有一定的调整：

1）液体润湿面材料与不同工质的兼容性存在差异，更换工质后需要重新开展材料与工质间的兼容性测试验证，以保证长期应用可靠性。

2）CDU：需要对补液装置改进，避免补液过程空气中的水分或杂质进入液冷系统中，引起非水工质的水解产生酸性物质，导致腐蚀风险问题。

3）漏液检测：二次侧冷却液为不导电液体，因此传统导电型漏液检测传感器不再适用，需要更换为光电式、电容式、浮子式漏液检测方式。针对氟碳类工质，因其挥发性较强，泄漏后有一定的气态工质产生，可以采用吸气式漏氟检测仪器。

3.2 全液冷冷板

传统冷板式液冷通常只覆盖CPU、GPU等个别高功耗芯片，设备节点或整机柜液冷占比通常在60%~80%之间，存

在液冷占比低、节能收益不显著的问题。为此，业内已经开始布局全液冷冷板技术，即通过液冷板为设备内的所有发热器件进行散热。

以通算服务器产品为例，液冷板覆盖CPU、内存、硬盘、电源等，95%以上的热量通过液冷板带走，剩余约5%的热量通过设备节点内风液换热器中的冷却液带走，进而实现100%液冷。与传统的单相冷板式液冷相比，全液冷冷板技术具有更低的系统能耗，PUE可低至1.1，能够有效降低数据中心的运营成本。

全液冷冷板虽然可以大幅提升液冷占比，提升节能效果，但涉及液冷部件较多，液冷系统相对复杂，需要专业的维护人员进行操作和维修，同时内存、硬盘等可插拔部件的应用可靠性仍有待提升。从长期收益来看，全液冷冷板技术得益于其高效的散热性能及更低的能耗，在数据中心领域会有更广泛的应用。

3.3 单相浸没强化散热

单相浸没液冷液体流速低，使系统解热能力受限。在当下智算如火如荼发展的过程中，高功耗、高热流密度的CPU/GPU散热需求，驱动人们不断探索散热强化的创新路线，如引入主动驱动力，调整系统架构，改善冷却液热物性等，以满足高功耗、高热流密度芯片的散热需求。

单相浸没液冷通过引入外部驱动部件，可以显著提升芯片局部区域的冷却液流速和湍流程度，实现较高的换热效率。例如，Submer和英特尔共同开发了一款强制对流散热器，其通过在翅片散热器前方加装风机，搭配限流器外壳，使在散热器鳍片区域的冷却液产生强制对流，提高了冷却液的换热效率，从而改善散热器的热性能[7]。

除了模块化设计的强制对流散热器方案，系统架构调整的散热模式也是浸没液冷发展方向之一。例如，中兴通讯与英特尔合作开发的浸没液冷架构强化方案，采用双回路设计，高功耗器件CPU/GPU等支持重力驱动强化散热方式，支持单节点散热能力2 000 W以上，CPU散热能力大于550 W[8]。

冷却液方面，目前单相浸没冷却液以碳氟类和油基工质为主，相比于水溶性液冷，虽然可以有效地解决绝缘性问题，但是仍存在粘度大、比热容低、导热能力差的缺点。为了提升介电液体的散热能力，纳米流体成为当下研究方向之一。纳米流体借助纳米颗粒的高导热系数和液体与颗粒之间的对流，可以显著提高导热系数和对流传热系数。虽然采用纳米流体可以有效提升换热性能，但是其稳定性差、制备难度大、生产成本高是实际应用中存在的主要问题，现阶段仍需要持续优化。

3.4 液冷智能温控技术

液冷系统的极致节能离不开管理层的优化调控。与风冷系统相比，液冷系统耦合性更强，系统控制点位更多、更复杂。传统的液冷系统调控逻辑或群控模式无法匹配业务和负载率变化进行主动调控，在一定程度上存在冷量浪费的问题。现阶段的AI调优测试主要基于数据模型，通过对历史数据的深度学习、强化学习等，仅利用有限场景下的纯数据样本，数据成本高，历史数据依赖性强，训练周期长，且不具有可解释性，容易反逻辑控制，在极端工况下可靠性低。

为了提高液冷系统温控策略的节能效果及运行稳定性，人们提出了“数据+机理”的双驱AI技术。该技术将AI与传统暖通热力学模型相结合，构建机理和数据融合驱动的系统热力学模型，并针对机理模型中难以建立“白箱”模型的部分，可以利用采集数据构建数据模型来解决，也可以利用数据驱动方法对机理模型中的参数进行优化。双驱AI控制策略遵循热学原理，脱离纯数据依赖，避免反逻辑，具有更高可靠性、更优节能效果，能够通过对两种预测模型取长补短，最大程度提高预测的准确性，使计算复杂度及成本显著降低。

在具体应用中，需要将尽可能地将影响液冷系统节能与运行稳定性的因素纳入数据中心基础设施管理（DCIM）监管和调控中，通过双驱模型对数据中心建立多输入和输出间的拟合关系，使各工况点均具有可预测性。融合机理模型和数据模型的双驱动仿真系统，借助可视化平台开发，可建立数据中心系统的数字孪生预测模型。液冷系统基于该模型不仅能实现极佳的节能温控策略，还能针对极端场景提前制定可能的风险场景应对策略，提升运维人员的响应效率和数据中心的运行可靠性[9]。

3.5 低成本液冷系统

与传统风冷系统相比，液冷技术应用存在初期投资成本高的问题，这影响了液冷技术的规模应用与推广。此外，液冷物料本身也需要进一步研究。原材料和加工成本较高，需要引入新材料或新工艺以进一步降低成本[10]。基于此，中兴通讯开发了低成本液冷系统，通过引入高可靠、低成本材料，改善工艺条件，使液冷数据中心投资成本综合降本15%以上。引入的材料包含铝合金冷板、高分子材料等。其中，高分子材料包括高分子工程管网、高分子分液器、高分子流体连接器等。

1）铝合金冷板：液冷板散热底板由铜材更换为铝材。冷板上盖板等非散热接触面材料采用高分子材料，并通过注

塑成型，降低了成本。同时，液冷板取消焊接密封工艺，采用胶圈密封的方式，节省了焊接费用。

2）高分子材料应用：工程管网、分液器等由不锈钢材料更换为高分子材料，且一体式注塑成型，工艺成本低，且所选材料经过兼容性测试验证，应用可靠性高。

3.6 芯片级液冷

芯片制程工艺向更小尺寸发展，芯片功耗和热流密度不断攀升，加之2.5D/3D封装和异构芯片的快速发展，使得芯片内热阻占比越来越大。当前芯片散热主要考虑导热界面材料（TIM）和外部系统散热技术两个方面，但仍无法解决芯片内热阻大的问题。未来随着各种新型封装形式的演进，外部液冷散热方案将难以满足超高功率密度芯片的散热需求。液冷散热方案将深入到芯片内部，从热源根本上解决散热问题。这种散热技术称为芯片级液冷技术。

芯片级液冷沿用冷板式液冷架构，所不同的是其将微尺度流道（微米级通道宽度）刻蚀在芯片内部，液体工质直接从芯片内部带走热量，大大降低芯片内热阻或者界面热阻，同时可解决多Die堆叠引起的散热问题，使散热能力得到极大提升，并可满足超高散热需求。从1981年开始，全球陆续有一些高校、科研机构和芯片厂商已经布局芯片级液冷散热技术研究，包括对微尺度液冷基础原理的研究、微尺度（硅基）流道加工工艺的探索改进、先进微尺度流道设计方案的研究等。按芯片与液冷微通道的耦合形态，芯片级液冷又可分为分体式（含TIM）和一体式（无TIM）两种，预计均可满足300 W/cm^{2}以上的散热需求。但由于相关技术成熟度还较低，目前业内还暂无应用案例。

4 结束语

在“数字经济”和“双碳”的大背景下，不断提升的芯片热流密度和更严苛的设备能耗设计要求，成为数据中心制冷技术不断演进的两大重要驱动力。液冷技术具有低能耗、高散热、低噪声、低TCO等优势，是解决芯片散热问题、打造绿色低碳数据中心的关键技术。

参考文献

[1] 工业和信息化部. 新型数据中心发展三年行动计划（2021—2023年）[R]. 2021

[2] 国家发展改革委, 中央网信办, 工业和信息化部, 等. 贯彻落实碳达峰碳中和目标要求 推动数据中心和5G等新型基础设施绿色高质量发展实施方案 [R]. 2021

[3] MATSUOKA M, MATSUDA K, KUBO H. Liquid immersion cooling technology with natural convection in data center [C]//Proceedings of IEEE 6th International Conference on Cloud Networking (CloudNet). IEEE, 2017: 1–7. DOI: 10.1109/CloudNet.2017.8071539

[4] 科智咨询, 中国信息通信研究院. 中国液冷数据中心市场深度研究报告 [R]. 2021

[5] HONDA H, TAKAMATSU H, WEI J J. Effect of the size of micro-pin-fin on boiling heat transfer from silicon chips immersed in FC-72 [J]. Transactions of the Japan society of mechanical engineers series b, 2002, 68(672): 2327–2332. DOI: 10.1299/kikaib.68.2327

[6] SARANGI S, MCAFEE E D, DAMM D G, et al. Single-phase immersion cooling performance in intel servers with immersion influenced heatsink design [C]//Proceedings of 38th Semiconductor Thermal Measurement, Modeling & Management Symposium (SEMI-THERM). IEEE, 2022: 1–5

[7] 英特尔与Submers合作推出浸没式液冷系统可为1000 W以上CPU散热 [EB/OL]. (2023-10-18) [2024-10-15]. https://www.expreview.com/90550.html

[8] 中兴通讯. 中兴通讯浸没式服务器IceTank, 重塑未来绿色数据中心 [EB/OL]. (2024-06-05) [2024-10-15]. https://www.zte.com.cn/china/about/news/20240625c2.html

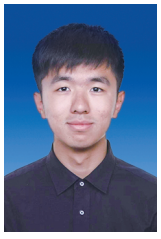
[9] 开放数据中心委员会. 数据中心暖通系统AI节能开放架构白皮书 [R]. 2023

[10] 开放数据中心委员会. 冷板液冷标准化及技术优化白皮书[R]. 2023

作 者 简 介

严劲，中国电信智能云网调度运营中心云室主任；长期从事云计算、数据中心的运行维护、性能调优、技术研究等工作，组织实施过多项云计算、数据中心领域的运营优化和业务保障重大项目，在云计算运营风险治理、架构优化等方面拥有丰富经验。

景焕强，中兴通讯股份有限公司制造工程研究院院长、硬件工程研发中心负责人；具有丰富的可靠性理论和实践经验，负责中兴通讯液冷团队一体化运作，推进液冷技术实现商用突破以及关键技术的差异化创新。



张子骜，中国电信集团云网调度运营中心项目经理，工程师；研究方向为算力网络、云计算及运维、人工智能、区块链。

刘帆，中兴通讯股份有限公司热设计技术总工、青年领军人才；研究方向为ICT设备先进散热技术预研、机房热管理技术预研，具有丰富的热设计经验与扎实的技术能力；申请30余项，发表论文5篇。

基于通信扩展定义的语义通信三层架构

Semantic Communication Three-Layer Architecture Based on Extended Definition of Communication

张黎明/ZHANG Liming

（国家发展改革委创新驱动发展中心（数字经济研究发展中心），中国 北京 100038）
(Center for Innovation-Driven Development, P. R. China (Center for Digital Economy Research and Development, NDRC, P. R. China), Beijing 100038, China)

DOI：10.12142/ZTETJ.202406014
网络出版地址：http://kns.cnki.net/kcms/detail/34.1228.TN.20241126.1111.002.html
网络出版日期：2024-11-26
收稿日期：2024-10-25

摘要：人工智能的快速发展为WEAVER与SHANNON所设想的通信扩展定义提供了技术实现基础。基于该通信扩展定义，提出了语义通信三层架构，将语义信息真正融入通信系统设计。发送端可以根据信息的语义重要性进行差异化编码和传输，接收端能够恢复发送端传输数据的语义，并根据语义重要性完成通信的语用目的。进一步地，基于深度学习实现的语义通信三层架构，优化了信息传递的有效性和可靠性，可以更好地满足未来6G新兴应用场景，符合国家战略发展需求。研究了语义通信三层架构搭建方法，系统总结了其中需要面对的挑战与需要实现的关键技术，并探讨了其在物联网、人机交互和智能制造中的广泛应用前景。

关键词：人工智能；通信扩展定义；语义；语用；差异化编码

Abstract: The rapid development of artificial intelligence has provided a technological foundation for the extended definition of communication envisioned by Weaver and Shannon. Based on this extended definition of communication, this paper proposes a three-layer semantic communication architecture that truly integrates semantic information into communication system design. The sender can perform differentiated encoding and transmission based on the semantic importance of the information, while the receiver can recover the semantics of the data transmitted by the sender and accomplish the communication's pragmatic purposes according to semantic importance. Furthermore, this three-layer architecture of semantic communication, implemented using deep learning, optimizes the effectiveness and reliability of information transmission, better meeting the emerging application scenarios of future 6G networks and aligning with national strategic development needs. This paper explores the methods for constructing the three-layer semantic communication architecture, systematically summarizes the challenges faced and the key technologies that need to be implemented, and discusses its wide application prospects in the Internet of Things, human-computer interaction, and intelligent manufacturing.

Keywords: artificial intelligence; extended definition of communication; semantics; objective; differentiated encoding

引用格式：张黎明. 基于通信扩展定义的语义通信三层架构 [J]. 中兴通讯技术, 2024, 30(6): 92-99. DOI: 10.12142/ZTETJ.202406014
Citation: ZHANG L M. Semantic communication three-layer architecture based on extended definition of communication [J]. ZTE technology journal, 2024, 30(6): 92-99. DOI: 10.12142/ZTETJ.202406014

1 6G与语义通信

信息技术应用场景的不断扩展，例如沉浸式通信、工业互联网、生成式人工智能等，不仅对数据传输需求产生了爆炸式增长，而且对6G通信技术的多个关键性能指标也提出了更高的要求，包括峰值速率、传输时延、流量密度和定位精度等。进一步地，6G也将不仅仅局限于网络容量和传输速率等指标的提升，它还需要为大规模物联网提供支持，连接数十亿设备，涵盖智能家居、智慧城市以及工业物联网等多个领域。

随着5G中各项技术的不断发展，基于SHANNON信息论[1]的各通信模块几乎达到了各自的理论极限，技术体系日益复杂，基站能耗也日益增加。如果6G继续沿用堆叠资源的方式实现，将不可避免地造成大量浪费，而这不符合国家可持续发展的战略方针要求。此外，中国在芯片技术方面仍

存在短板，积极研究有别于传统通信方式的6G新型通信芯片技术，有助于国家在芯片研究方面锻造新长板，实现弯道超车，推进网络强国的国家战略方针。因此，在已有资源条件和当前芯片制程下，探索新的通信方式成为实现6G的必然选择。

事实上，通信底层理论并非只有大众熟知的经典信息论。1949年，WEAVER和SHANNON对通信做了进一步阐述[2]，以扩展Shannon原始对通信定义中的工程性假设，即“These semantic aspects of communication are irrelevant to the engineering problem.”在新的定义中，通信被系统描述为解决三个层面的问题：

LEVEL A：通信符号能多精确地传输？

LEVEL B：传输符号能多准确地传达所期望的含义？

LEVEL C：接收含义能多有效地以期望的方式影响行为？

一般将其翻译为通信的语法问题、语义问题以及语用问题。语法问题关注传输符号的准确性，语义问题关注传输符号是否准确表达了其背后的含义，语用问题则关注接收方接收的信息是否能按预期完成通信目的。基于SHANNON经典信息论的传统通信很好地解决了语法问题，但是面对后两个问题，传统通信将导致大量冗余信息的传输，并非该通信扩展定义的最佳实现手段。

针对WEAVER和SHANNON对通信的延伸定义，也为了发展下一代无线通信技术，实现国家战略方针，本文提出了能够解决语法、语义、语用问题的通信系统三层架构，简称语义通信三层架构。该架构基于传统通信方式，面向语义问题，通过深度学习技术提取传输符号的语义特征，对传输数据进行精简，从而大幅提升通信效率。进一步地，针对特定通信任务，基于生成式人工智能技术，衡量所提取的语义特征对智能任务完成的不同贡献度，即语义重要性，生成带有语义重要性评分的多个语义特征向量（SFV），从而解决语用问题。

2 语义通信三层架构体系框架

本文建立了一种基于通信扩展定义的语义通信三层架构，该架构涵盖了WEAVER与SHANNON所述的语法问题、语义问题以及语用问题。基于传统通信框架，该架构依然分为信源、信道、信宿等模块，但是以提取和传输数据语义为核心。如图1所示，语义通信的三层架构主要由语法层、语义层和语用层组成。首先在语义层面，系统着重于如何精确传达数据所蕴含的内在含义；其次在语法层，系统侧重于如何准确传输通信符号，但这一过程受到语义层面通信的指导；最后基于所传输的符号语义，在语用层，系统针对SFV评分对接收到的数据语义进行处理，从而更好地完成通信智能化任务。此外，语义编码可以选择分离实现，也可以采用联合信源信道编码的方式实现，即编码器的压缩（信源编码）和纠错（信道编码）功能。这样既可以按照传统系统中的模块化方法实现，也可以依据联合源信道编码理论，采用集成化的实现方式。

需要说明的是，影响通信质量的因素主要包括语义噪声和信道噪声。其中信道噪声，是受传输设备和复杂传输环境等物理因素影响，使得传输数据失真的噪声，它是语法层面的噪声，在传统通信研究中占据重要地位。因此，语义通信三层架构中的语法层，要研究信道噪声如何影响语义传输。

2.1 语法层

在语义通信三层架构中，虽然以语义层为核心，但符号或比特依然是信息传输的物理承载，因此语法层的主要任务为打通数据统计特征与数据语义特征之间的关系，完成承载语义特征的数据传输。根据文献[3–5]所述，语义特征实际上是通过精简数据冗余得到的，而精简数据的过程也是语义编码过程和语法层传输过程。此外，虽然信道噪声会影响数据传输的准确度，但是信道噪声并不总会影响语义传输的准确度，即语义对信道噪声存在一定的容忍限度。因此，语法层需要研究的主要问题为，噪声是如何影响数据与其语义关系的。

根据最新的语义信息熵研究[3]，引入同义映射的概念即可描述数据空间与语义空间的关系，并根据同义映射，可以定义出语义熵。其中对于高斯分布的语义熵公式为：

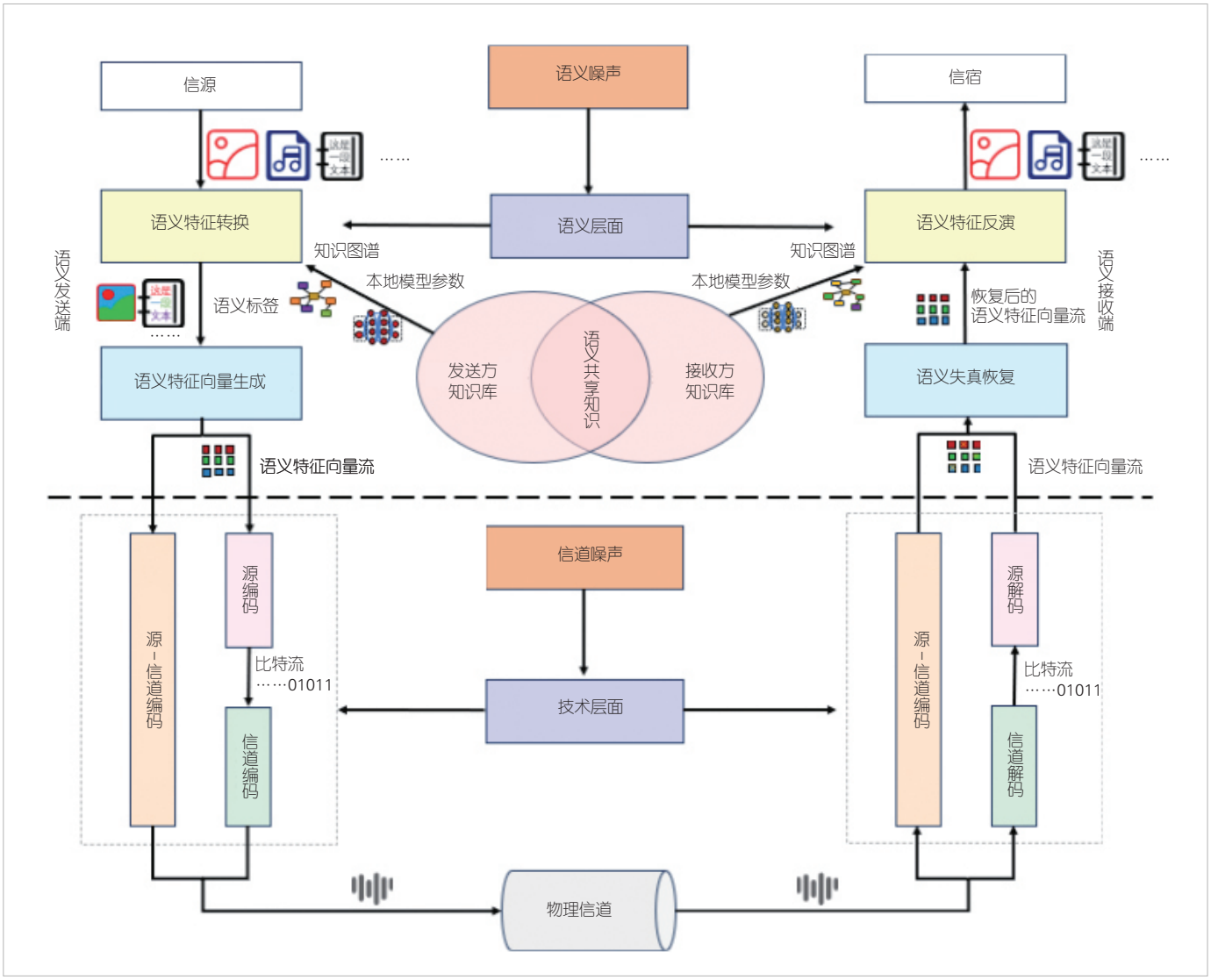
$$H_S = \frac{1}{2}\log\frac{2\pi e\sigma^2}{S^2}, \tag{1}$$

其中，S表示同义映射的长度，对于离散情况，可以简单理解为一个语义基对应的数据数。σ^2表示数据分布的方差。由于存在同义映射，因此根据数据统计分布计算的熵，小于根据数据语义计算的熵。进一步地，假设信道噪声为可加高斯白噪声，可以根据同义映射得到语义传输的信道容量：

$$C_S = B\log\left[S^4\left(1+\frac{P}{N_0B}\right)\right], \tag{2}$$

其中，P表示信号功率，N_0表示信道噪声功率谱密度，B表示可用带宽。

根据上述研究，可以发现语义通信对信道噪声具有一定的容忍限度，因此可以指导语义通信三层架构中的语法层设计。语义通信的主要目的为，能够保真传输表达语义的符号或者比特。此外，针对不同模块的详细设计思路，还需要发



▲图1　基于WEAVER与SHANNON对通信扩展定义的语义通信三层架构

展进一步的语义编译码理论，可以参考3.2节中的最新研究。

2.2 语义层

由于语义通信三层架构围绕语义理解，基于语法传输，最终实现通信的语用目的，因此语义层是语义通信三层架构的核心层。语义层的相关技术研究较多，因此这里只给出具体结构模块，关键技术实现可以参考3.2节。

• 信源：负责产生需传递给接收者的数据。这些源数据可以是文本、语音或音频、图像、视频等形式，抑或是这些形式的组合，例如集成了视频和相应音频的多媒体数据。甚至在未来，还可以是脑电波等能够表征人类或者智能意识的数据。

• 语义特征转换模块：该模块对信源产生的数据进行处理。一方面，该模块需要依据本地数据库中的知识图谱和训练数据集等信息，提取数据的语义特征，并根据这些特征为源数据产生相应的语义标签。另一方面，该模块需要考虑语法层中分析的第4部分信道噪声，尽可能产生与之匹配的语义编码。关键技术实现放在3.2节中。

• 语义特征向量生成模块：该模块可以协助语用层实现智能化通信的语用目的，进一步提高通信效率。针对特定任务，不同的语义特征对智能任务的完成具有不同的贡献度，即语义重要性。此模块利用本地知识库中的语义重要性评分，评估语义标签中每个语义特征的价值，并生成带有语义重要性评分的多个SFV。该模块的功能与通信目的紧密相关。在人际通信中，人们期望获取全面的信息，此时系统会采用显著性检测等技术来确定SFV的重要性分数；而在机器

间通信时，如人脸识别任务中，机器无需获取全部信息，例如可以忽略背景。此时可运用度加权类激活映射等方法生成SFV的重要性分数，以进一步减少传输信息量，提升通信的效率与准确度。

• 语义失真恢复模块：该模块可以补足被第2.2和第3部分噪声影响的语义信息。利用发送端提供的相关辅助信息，对受损的语义向量流进行修复。这些辅助信息可能包括编码后的语义标签、传输错误掩码图等含有发送端源数据特征的信息。其关键技术实现可以参考3.2节中基于生成对抗网络（GAN）的语义失真恢复方法。

• 语义特征反演模块：执行语义特征转换的逆过程。利用恢复后的语义特征向量进行语义特征融合，以重构源数据或直接驱动下游的机器智能任务。

• 信宿：指源数据传输的最终对象，可以是个人或者物体。

• 本地知识库：隐性存在于发送端和接收端，为系统提供先验知识。该模块不一定是物理实体，例如，可以是用于训练深度神经网络的数据集，这些网络用于生成语义指导和源信道编码，使得网络参数成为知识的一种体现；也可以是知识图谱，通过使用知识图谱实现语义压缩，仅传输最相关的语义信息，减少传输数据量的同时不丢失关键意义。一般而言，由于发送端与接收端的本地知识库通常存在差异，所以需要定期通过云端服务器进行数据共享，以实现知识库的同步，其关键技术实现可以参考3.2节中的研究。

2.3 语用层

语用层是语义通信三层架构的最上层，也是WEAVER和SHANNON通信定义的论述中，最终要实现的一层。该层的功能主要依靠SFV进行差错编码实现，其相关技术也在3.2节中进行了论述，主要包括如下结构：

• 源解码器：该模块根据语义重要性分数，对分数高于设定阈值的SFV进行编码，并对每个SFV实施压缩处理，从而生成比特数量较少的数据流。

• 信道编码器：通过引入卷积码、涡轮码等冗余信息，增强数据在传输过程中的抗信道噪声和干扰能力，确保数据的可靠传输。

• 源–信道编码器：对每个SFV进行处理，以生成适合在信道中传输的信号。该模块的设计受到语义重要性分数的指导，采用联合信源信道编码策略，在设计编码方案时兼顾源编码与信道编码的需求，对语义噪声和信道噪声进行平衡，旨在减少传输信息量的同时，降低语义失真恢复的复杂性，提高通信的有效性与可靠性。

• 信道：作为从发送端到接收端传输信号的物理介质，其作用与经典通信系统中的信道相同。

• 源–信道解码器：执行信源–信道编码器的逆过程，以重建SFV。

• 源解码器：对压缩后的比特流执行解压缩操作，恢复出原始的SFV。

• 信道解码器：进行信道编码器的逆操作，将接收到的信道信号转换为相应的比特流。

3 语义通信三层架构关键技术

语义通信并非全新领域，已经存在大量理论、实验与系统搭建研究方法和思路。因此，本文在第二节中提出的语义通信三层架构，其中大量关键技术，可以借鉴一些最新研究实现。本节将概述几个主要挑战，并介绍针对这些问题的最新研究成果。

3.1 面临的挑战

1） 理论体系尚待完备

虽然WEAVER和SHANNON提出语义通信是经典通信的进一步延伸，但与基于经典信息论的传统编码相比，基于机器学习的语义通信系统在理论分析上依然缺乏系统的数学理论支撑。这主要是由于语义基难以定义导致的，只有定义出语义基，才能定义出语义空间，也才能在语义空间中寻找合适的范数公式与距离衡量公式。有了这两种度量才有类似于概率空间中的熵出现，即语义空间中的信息度量方法。语义度量方法能够帮助我们对语义通信三层架构中的语义特征变换和语义失真校正进行准确界定和评价，指导其中语法层的具体实现。此外，我们只有研究信道在语义空间中的表征，即研究语义度量下的信道噪声，才能对语义通信系统的传输速率有明确的界限认知（像SHANNON限那样），也才能指导语义通信三层架构中语义层和语用层的实现。

2） 异质异构知识库

在语义通信过程中，发送端与接收端均配备本地知识库。但是，不同端侧的知识库往往存在差异。这些差异不仅体现在知识库内存放的先验知识可能不同，而且还体现在这些先验知识的组织架构不尽相同。尽管通过共享可以增进知识库的一致性，但这一过程既耗时间又耗资源。这是因为知识库的共享依赖于发送端与接收端之间的有效通信。知识库需要不断扩展和更新。如同人类学习的过程一样，知识库共享过程更加漫长且复杂。因此，如何在知识库不一致的情况下进行通信、共享和推理语义信息，是实现语义通信三层架构面临的另一大挑战。

3） 语义失真的精准恢复

在语义通信中，尤其是在机器间的通信场景中，信源编码和信道编码会根据语义重要性分数对传输信息进行筛选，仅保留较为重要的部分。这种做法虽减轻了信道传输的负担，但也增加了语义噪声，为语义信息失真的恢复带来了难度。语义失真一旦太大，三层架构中的语用层功能即会受到极大的影响。

4） 语义系统评价指标尚待完备

在语义通信三层架构中，不同类型的信源数据采用不同的深度学习模型，产生的语义标记信息各异，因此，为不同类型的信源数据设计适用于语义通信系统的评估指标显得尤为重要。此外，还需要建立一个通用的性能指标，如同传统通信系统中的符号错误率（SER）或比特错误率（BER），以衡量不同语义编译码模型的性能。

3.2 关键技术

1） 语义通信基础理论研究

文献[4]对语义通信的数学理论基础进行了进一步探讨，通过详细界定一系列基本的语义相关概念和语义编码方案的数学表达，构建了语义语言系统。此外，还提出了针对语义编码模型的统一语义平均失真评估公式：

$$D_{U,Q}=\sum_{w,s,\hat{s},\hat{w}}p(w)u(s|w)c(\hat{s}|s)q(\hat{w}|\hat{s})d(w,\hat{w}),\tag{3}$$

其中，w、$\hat{w}\in W$假设为发送端和接收端的传输和重构意义。$p(w)$是源信息w的概率分布，描述了不同意义w被传输的可能性。$u(s|w)$是编码方案U中，给定意义w编码为符号s的条件概率，反映了编码过程中符号s是如何根据意义w被选取的。$c(\hat{s}|s)$是信道模型中，发送符号s被接收为符号$\hat{s}$的条件概率，描述了信道在传输过程中引入的噪声或干扰影响。$q(\hat{w}|\hat{s})$是解码过程中的条件概率，给定接收到的符号$\hat{s}$重构出意义$\hat{w}$的概率，展示了解码器将符号转化回意义的能力。$d(w,\hat{w})$为语义失真度量，表示原始意义w和重构意义$\hat{w}$之间的差异程度，其映射为从$W\times W$到非负实数集，失真值越小，表示传输效果越好。此外，还定义出了语义平均成本计算公式：

$$L_U=\sum_{w,s}p(w)u(s|w)l(s),\tag{4}$$

其中，$l(s)$是信息s的成本函数，表示传输信息的代价。

基于上面的两个定义，文献[4]建立了一套优化语义编码与解码策略的方法，在数学上严格描述了不同方案下的语义失真和通信成本，形成了一套较为完整的语义通信理论框架。然而，该理论在实际应用中尚未充分考虑多模态以及复杂约束环境下的语义通信问题，而且也没有解决最重要的语义空间与数据空间的实际编码映射问题。这些问题还需要进一步研究。

2） 异质异构知识库同步

在语义通信中，通信双方拥有同样的知识库并不现实。知识库的异质异构性将影响通信性能。因此，文献[6]提出了一种高效的知识库同步机制和两种算法，分别为同步信号驱动模型估计（SSME）和数据模型迭代优化方法（DAMIO），旨在缓解由双方知识库异构性（尤其是数据分布差异）所引起的语义通信系统性能下降问题。

在语义通信三层架构中，如果存在知识库异构的问题，语义失真可以被建模为原信息x和恢复信息$\hat{x}$之间的不相似性，表示为：

$$\tilde{e}_{\theta_s,d_{\theta_r}}=F(x,\hat{x})=\left\|x-\hat{x}\right\|_2^2,\tag{5}$$

其中，θ_s、θ_r分别表示不同的知识库参数，d_{θ_r}表示以θ_r为代表的知识库进行语义解码的解码器参数。此时，具有相同知识库的语义失真可以表示为$\tilde{e}_{\theta_s,d_{\theta_s}}$，比不同知识库的语义失真要小，因此可以将由知识库异质导致的语义通信系统性能水平下降（PDL）表示为：

$$\Delta PDL_{\theta_s,d_{\theta_r}}=\tilde{e}_{\theta_s,d_{\theta_r}}[x,\hat{x}]-\tilde{e}_{\theta_s,d_{\theta_s}}[x,\hat{x}]。\tag{6}$$

进一步地，可以将需要优化的参数表示为：

$$\arg\min_{\theta_r}\Delta PDL_{\theta_s,d_{\theta_r}}=\arg\min_{\theta_r}\{\tilde{e}_{\theta_s,d_{\theta_r}}[x,\hat{x}]-\tilde{e}_{\theta_s,d_{\theta_s}}[x,\hat{x}]\}。\tag{7}$$

如果要减少语义失真，接收端必须尽力与发送端的参数保持一致（为联合优化参数），但是如果直接传输参数进行同步，需求的数据传输量过大，因此可以分三步优化：

• 初始化：使用上述同步算法的前提在于，发送端与接收端具有一部分公共知识库，或者说是公共训练集D，对于公共训练集中数据，语义通信的失真程度小于一个任意小的值ε：

$$\xi_{\tilde{e}_{\theta_s,d_{\theta_s}}}[D,\hat{D}]=\sum_{d_u\in D}F_{\tilde{e}_{\theta_s,d_{\theta_s}}}(d_u,\hat{d}_u)<\varepsilon。\tag{8}$$

• 同步：当发送端与接收端的知识库不相同时（语义译码器参数与编码器参数非联合训练），发送端可以发送公共训练集D的语义特征$Z=\{z_i;z_i=e_{\theta_s}(d_i)\}$，接收端基于语义特征$Z$使用SSME或者DAMIO算法优化译码器，实现知识库同步。

• 通信：第二步中的语义特征Z可以视为知识库同步信

号，该同步信号可以定期发送，也可以在发送端数据分布出现显著变化时立即发送，得到同步信号后接收端会即时更新语义译码器参数，之后可以正常进行语义通信。

该机制将异构异质知识库问题转化为发送方和接收方语义编译码模型的估计问题，可以用于语义通信三层架构中语义层搭建。

3） 语义信息失真校正

文献[7]探索了结合GAN的语义失真恢复方法，并提出了一个基于文本传输且不考虑信道状态信息的语义通信框架（Ti–GSC）。该框架包含一个自动编码器模块和一个基于GAN的信号失真抑制模块。信号失真抑制模块利用GAN的生成能力，通过学习接收信号与传输信号之间的语义映射，生成在语法和语义上与传输信号相似的信号，从而在接收端实现更准确的数据解码。两个Transformer模块分别构成了自编码器模块的编码器和译码器，在通信中为语义联合编译码器。U–Net模块与Discriminator模块则构成了生成对抗网络的生成器与鉴别器，其中生成器在通信中负责辅助解码器进行语义解码，被称为基于GAN的非CSI信号失真抑制模块（GSDSM），鉴别器则用于压制语义失真。

进一步地，该架构的关键在于如何构建GAN的损失函数，判断传输文本的语义失真程度，从而进行失真恢复。文中通过引入句法损失（低维失真）和句义损失（高维失真）两个新的损失项，解决了这个问题：

$$R(X,\bar{Y}) = E[D_{\text{sytc}}(\bar{Y},Y)] + E[D_{\text{smtc}}(\bar{Y},Y)], \tag{9}$$

其中，X为发送端的语义编码，Y为通过信道传输后，受噪声影响的语义编码，$\bar{Y}$为经过GSDSM校正后的语义编码，$D_{\text{sytc}}(\bar{Y},Y)$为句法损失，用L2范数表示：

$$E[D_{\text{sytc}}(\cdot)] = E\left[\left\| X - G_{\text{nn}}^{\text{gnrt}}(Y|\theta_g)\right\|_2^2\right], \tag{10}$$

其中，$G_{\text{nn}}^{\text{gnrt}}(Y|\theta_g)$表示GSDSM模块函数，输出为$\bar{Y}$，$\theta_g$为表示该函数中的可训练参数。此外，$D_{\text{smtc}}(\bar{Y},Y)$表示句义损失，同样用L2范数表示：

$$E[D_{\text{smtc}}(\cdot)] = E\left[\left\| f(X) - f(G_{\text{nn}}^{\text{gnrt}}(Y|\theta_g))\right\|_2^2\right], \tag{11}$$

其中$f(\cdot)$表示语义编码器中间层函数，即句义损失是通过衡量文本在语义编码器中间隐藏层输出的向量L2范数得到的。

值得一提的是，上述研究中使用GAN衡量语义距离的方法，如果将上述鉴别器目标改为可达成语用目的的文本，则可以被迁移到本文所提语义通信三层架构中，作为实现语用层的核心技术。

4） 语义信息性能指标设计研究

在文本传输任务中，传统的语义通信系统[8–9]通常采用平均语义失真作为性能评价指标，而基于机器学习的语义通信系统[10–11]则普遍使用BLEU作为评价指标。此外，比较常用的文本评价指标为句子相似度：

$$match(M,\hat{M}) = \frac{B_{\Phi}(M)\cdot B_{\Phi}(\hat{M})^T}{\left\|B_{\Phi}(M)\right\|\left\|B_{\Phi}(\hat{M})\right\|}, \tag{12}$$

其中，B_{Φ}表示使用预训练模型，比如BERT将原始文本与接收文本嵌入为语义特征向量。最近的一些语义通信系统还提出了“语义相似性的上尾概率”[12]作为评价指标，通过计算接收语句与发送语句的语义相似性大于或等于某一阈值的概率，来评估语义通信系统在噪声干扰条件下的可靠性。虽然我们可以借助该指标在复杂的信道环境中量化语义通信系统的性能，但仍需要进行复杂的概率计算。此外，该指标还比较依赖于训练数据和模型的有效性。

对于图像传输任务，最常用的评估指标为峰值信噪比（PSNR），其计算方法如下：

$$PSNR = 10 \times \log_{10}\frac{L^2}{MSE}, \tag{13}$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{x}_i)^2, \tag{14}$$

其中，L表示图像中的像素最大值，一般为255。此外，最新的研究[13]提出了一种新的度量指标ViTScore。与传统的度量方法，如PSNR、多尺度结构相似性（MS–SSIM）、学习感知图像块相似度（LPIPS）等相比，ViTScore能够更准确地衡量图像的语义相似性，而非数据相同性。ViTScore基于预训练的视觉变换器模型，具有对称性、有界性和归一化等优点。

在语音传输任务中，可以运用自然语言处理（NLP）技术将语音转换为文本，然后以文本传输的方式在语义通信中实现。此外，考虑到语音还包含了发言人的情感信息，这些信息通过音调和音量的变化表现出来，因此仅仅进行语音翻译并按文本传输无法囊括语音中的所有语义信息。文献[14]提出可以采用信号失真比（SDR）和客观语音质量评估（PESQ）指标，来反映针对语音的语义通信传输的准确性。其中，SDR用于衡量语音信号在传输过程中的失真程度，而PESQ则综合考虑了情感信息。

目前在单模态语义信息传输的性能指标方面已有大量研究，但针对多模态通用语义通信系统的研究仍然相对较少。一种研究思路是将多模态信息统一为一种形式，并使用统一

的性能指标进行评估。文献[15]提出了一个语义信息的表示框架语义基（Seb）。它提供了一种模块化和高度抽象的方法来表示语义信息，从而提高了语义通信的效率。Seb可以包含与用户意图相关的背景知识、意图知识映射机制、语义元素提取和表达。通过这种表示框架，Seb将信息的多模态特性转化为语义元素，实现了更高效的信息传递。

但是，一方面，由于语义的复杂性和模糊性，上述指标依然难以覆盖所有的语义信息，无法组成一个完备的语义通信系统评价体系；另一方面，语义这个概念虽然是在通信领域最早被提出，但在人工智能尤其深度学习领域却研究得更多。因此，为了进一步研究能够完备衡量语义通信的性能指标，还需要从深度学习中引入更多语义衡量指标，这些指标如表1所示。

4 语义通信三层架构应用前景

相较于传统通信，语义通信三层架构显著简化了机器间的通信过程，为其在物联网、人机交互以及智能制造等领域的应用开辟了广阔前景。

4.1 物联网

在5G网络环境下，物联网（IoT）设备在天气监测、地理信息、智慧城市和家庭自动化等多种数据监控应用中扮演着关键角色。进一步地，虚拟现实（VR）/增强现实（AR）眼镜、无人机和传感器等智能设备的普及，需要IoT网络提供更高级的功能。IoT设备需要感知周围环境并实时将状态信息上传至云中心服务器进行分析处理。因此，它们必须能够支持智能监控、数据处理、实时通信等复杂功能，而这些功能的实现严重依赖低延迟、高准确性的数据传输。

语义通信提取并传输数据的抽象语义特征，能够实现准确、实时的数据传输。然而，由于IoT设备的计算和存储能力受限，难以直接部署复杂的深度神经网络，本地知识库的规模受限。因此，如图2所示，可以基于语义通信三层架构，通过网络稀疏化、神经元量化、联邦学习和分布式学习等技术，部署云端知识库，将知识图谱、模型参数等训练结果统一分发给参与通信的IoT设备，从而大幅度提升数据的传输效率，同时降低了IoT设备的计算负担，为物联网的未来发展提供了强有力的技术支持。

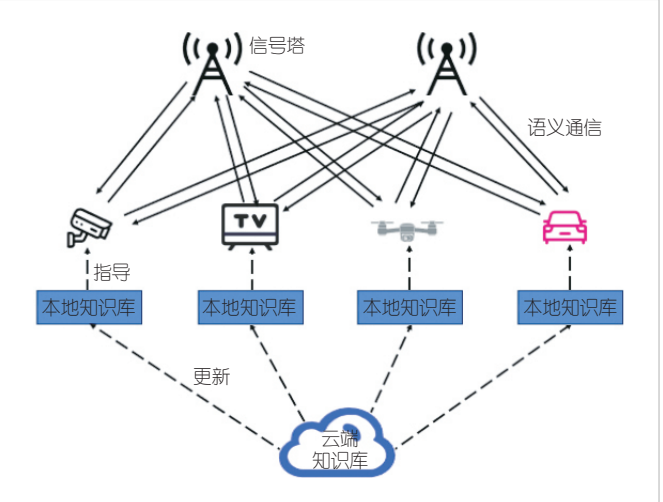
▼表1　语义通信三层架构可借鉴的深度学习领域语义级评价指标

深度学习中的语义级评价指标	
文本	平均语义失真
	词错误率（WER）
	N-gram
	双语评估替换（BLEU）
	基于共识的图像描述（CIDEr）
	基于BERT的相似性度量
语音	字符错误率（CER）
	信号失真比（SDR）
	语音质量的感知评估（PESQ）
	基于Fréche距离的语音深度学习度量（FDSD）
	基于Kernal距离的语音深度学习度量（KDSD）
图像	均方误差（MSE）
	峰值信噪比（PSNR）
	结构相似性指标（SSIM）
	Fréchet初始距离（FID）
	Kernel初始距离（KID）

4.2 人机交互

传统的人机交互主要依赖于预设指令和响应模式，信息传递多聚焦于数据层面。这样的交互方式往往使得机器在理解用户意图时不够准确，反馈也显得不够灵活，存在很大的局限性。例如，传统的语音助手大多只能基于关键词进行匹配，缺乏对用户指令的深层含义及其上下文关联的深入理解。相较之下，基于语义通信三层架构，可以解析信息的语义特征，同时结合人的通信目标，得到语义重要性分数。通过仅传输与人交互相关的关键语义特征信息，不仅简化了通信过程，还让机器能够更准确地理解用户意图，使语音助手和对话系统能够提供更加精确的回应。

特别地，当用户提出复杂问题时，基于语义通信三层架构搭建的系统可以依托上下文进行推理和理解，提炼出用户的核心需求，并据此提供更具针对性的回答，而非仅仅局限



▲图2　物联网中的语义通信

于简单的、模式化的响应。这种智能化的理解能力将会显著提升用户体验，使人机交互的智能化水平得到质的提升，让用户与机器之间的交流变得更加流畅和富有成效，互动过程更加自然和高效。

4.3 智能制造

在现代工业领域，智能制造正发挥着举足轻重的作用。随着6G和基于深度学习的通信技术等先进通信方式引入，智能制造正逐步实现智能化、高效化，并朝着节能环保的方向发展。

基于语义通信三层架构可以极大地提升机器间通信和人机交互的效率，进而增强智能制造的整体效能。在智能制造的实践中，语义通信技术为知识管理、自动化生产线的自配置，以及协作制造等应用提供了有力支持。通过在语义层面的数据处理和解释，生产流程得以优化，停机时间得以减少，生产效率可以得到显著提升。此外，语义通信系统还能够监控信息的语义特征，如机器状态、温度、湿度等关键参数，并能够将这些信息提取并上传至中央控制器，以进一步分析材料状态和产品质量，从而实现对生产过程的精细化管理。

5 结束语

在6G通信愿景中，语义通信技术展现出巨大的应用潜力，尤其是在物联网、人机交互和智能制造等领域。尽管相关研究已取得显著进展，但语义通信的实现仍面临诸多挑战。其中，基础理论尚未完全成熟，异构和异质的知识库存在复杂性，语义信息的失真恢复仍是一大难题。但是，随着人工智能技术的不断发展，语义通信的理论研究有望进一步完备，应用实践或许可以进一步完善。本文提出的语义通信三层架构，能够推动通信技术的进一步创新，满足未来信息服务日益增长的数据传输需求，促进科技进步和社会发展。通过不断攻克相应挑战，语义通信将在未来的通信网络中扮演更加重要的角色，为构建智能化社会提供坚实的技术支撑。

参考文献

[1] SHANNON C E. A mathematical theory of communication [J]. The bell system technical journal, 1948, 27(3): 379–423. DOI: 10.1002/j.1538–7305.1948.tb01338.x
[2] WARREN W. Recent contributions to the mathematical theory of communication [J]. ETC: a review of general semantics, 2017, 74 (1/2): 136–157
[3] NIU K, ZHANG P. A mathematical theory of semantic communication [EB/OL]. [2024–10–15]. http://arxiv. org/abs/2401.13387
[4] SHAO Y L, CAO Q, GÜNDÜZ D. A theory of semantic communication [J]. IEEE transactions on mobile computing, 2024, 23(12): 12211–12228. DOI: 10.1109/TMC.2024.3406375
[5] FENG Y L, XU J, LIANG C L, et al. Decoupling source and semantic encoding: an implementation study [J]. Electronics, 2023, 12(13): 2755. DOI: 10.3390/electronics12132755
[6] XU X D, BIAN Z Q, WANG B Z, et al. Synchronization mechanism: preliminary attempt to enable wireless semantic communication with heterogeneous knowledge bases [J]. IEEE communications letters, 2024, 28(8): 1815–1819. DOI: 10.1109/LCOMM.2024.3412811
[7] MAO J, XIONG K, LIU M, et al. A GAN–based semantic communication for text without CSI [J]. IEEE transactions on wireless communications, 2024, 23(10): 14498–14514. DOI: 10.1109/TWC.2024.3415363
[8] GULER B, YENER A. Semantic index assignment [C]//Proceedings of IEEE International Conference on Pervasive Computing and Communication Workshops (PERCOM WORKSHOPS). IEEE, 2014: 431–436. DOI: 10.1109/PerComW.2014.6815245
[9] GÜLER B, YENER A, SWAMI A. The semantic communication game [J]. IEEE transactions on cognitive communications and networking, 2018, 4(4): 787–802. DOI: 10.1109/TCCN.2018.2872596
[10] XIE H Q, QIN Z J, LI G Y, et al. Deep learning enabled semantic communication systems [J]. IEEE transactions on signal processing, 2021, 69: 2663–2675. DOI: 10.1109/TSP.2021.3071210
[11] XIE H Q, QIN Z J. A lite distributed semantic communication system for Internet of Things [J]. IEEE journal on selected areas in communications, 2021, 39(1): 142–153. DOI: 10.1109/JSAC.2020.3036968
[12] GETU T M, SAAD W, KADDOUM G, et al. Performance limits of a deep learning–enabled text semantic communication under interference [J]. IEEE transactions on wireless communications, 2024, 23(8): 10213–10228. DOI: 10.1109/TWC.2024.3370497
[13] ZHU T T, PENG B, LIANG J F, et al. How to evaluate semantic communications for images with ViTScore metric? [J]. IEEE transactions on cognitive communications and networking, 2024, 10(5): 1744–1758. DOI: 10.1109/TCCN.2024.3392803
[14] WENG Z Z, QIN Z J, LI G Y. Semantic communications for speech signals [C]//Proceedings of ICC 2021 – IEEE International Conference on Communications. IEEE, 2021: 1–6. DOI: 10.1109/ICC42927.2021.9500590
[15] ZHANG P, XU W J, GAO H, et al. Toward wisdom–evolutionary and primitive–concise 6G: a new paradigm of semantic communication networks [J]. Engineering, 2022, 8: 60–73. DOI: 10.1016/j.eng.2021.11.003

作 者 简 介

张黎明，国家发展改革委创新驱动发展中心（数字经济研究发展中心）高级工程师；主要研究方向为半导体、商业航天、人工智能、新型显示等新兴产业领域。

面向5G NR L2协议安全的自动化模糊测试技术

Automated Fuzzing Technology for Security of 5G NR L2 Protocol

钟宏/ZHONG Hong[1,2]，夏云浩/XIA Yunhao[1,3]，
张金鑫/ZHANG Jinxin[1,3]，马致原/MA Zhiyuan[1,3]
（1. 移动网络和移动多媒体技术国家重点实验室，中国 深圳 518055；
2. 深圳市中兴软件有限责任公司，中国 深圳 518057；
3. 南京中兴新软件有限责任公司，中国 南京 210012）
(1. The State Key Laboratory of Mobile Network and Mobile Multimedia Technology, Shenzhen 518055;
2. Shenzhen Zhongxing Software Company Limited, Shenzhen 518057, China;
3. Nanjing Zhongxing new Software Company Limited, Nanjing 210012, China)

DOI：10.12142/ZTETJ.202406015
网络出版地址：http://kns.cnki.net/kcms/detail/34.1228.TN.20240726.1711.006.html
网络出版日期：2024-07-29
收稿日期：2024-06-15

摘要：5G协议的安全性直接影响到5G通信系统能否正常提供服务，而新空口（NR）协议是其重要组成部分，因此对5G NR协议进行安全检测具有重要意义。提出一种基于模糊测试的5G NR协议漏洞检测自动化系统，针对媒体接入控制层（MAC）、无线链路控制层（RLC）和分组数据汇聚协议层（PDCP）的L2协议，分析协议特征来设计高效的数据变异策略，提高测试用例的有效性，实现多种工作模式适配以提高漏洞挖掘效率。接着，基于5G基站和移动终端设备，开发了一套原型系统用以评估本文所提方案的性能。实验数据显示，数据包处理时间能够满足5G业务时延性能要求，同时能够发现MAC、RLC和PDCP协议的多种安全漏洞，验证了所提方案可以有效提升测试数据包的合法性和漏洞挖掘的有效性。

关键词：5G NR；网络协议；漏洞挖掘；模糊测试

Abstract: New radio (NR) is an important part of the 5G protocol, and its security directly affects whether the 5G communication system can provide services properly. Therefore, it is of great significance to perform security testing on the 5G NR protocol. In order to efficiently perform security detection on medium access control (MAC), radio link control (RLC) and packet data convergence control (PDCP) of 5G NR L2 protocol, this paper proposes an automated system based on fuzzing technology. The proposed method designs efficient data mutation strategies by analyzing protocol characteristics to improve the effectiveness of test cases, and implements multiple working modes to improve the efficiency of vulnerability detection. Furthermore, in order to evaluate the performance of the proposed method, we implement a fuzzing prototype system based on 5G gNodeB and mobile terminal, and then conduct practical security detection on 5G NR protocol. Experimental results show that packet processing time of our proposed method can meet 5G latency requirements. In addition, various vulnerabilities in MAC, RLC, and PDCP are exposed in this paper which verifies that the proposed method can effectively improve the compliance of test data and the effectiveness of vulnerability detection.

Keywords: 5G NR; network protocol; vulnerability detection; fuzzing test

引用格式：钟宏, 夏云浩, 张金鑫, 等. 面向5G NR L2协议安全的自动化模糊测试技术 [J]. 中兴通讯技术, 2024, 30(6): 100-107. DOI: 10.12142/ZTETJ.202406015
Citation：ZHONG H, XIA Y H, ZHANG J X, et al. Automated fuzzing technology for security of 5G NR L2 protocol [J]. ZTE technology journal, 2024, 30(6): 100-107. DOI: 10.12142/ZTETJ.202406015

5G通信技术具备更高速率、更大连接、更低时延等技术优势，使得5G通信网络得到大规模部署和应用，在社会生活中发挥着重要作用。5G安全将直接影响到行业安全甚至是国家安全。其中，5G协议是保证5G通信系统能够正常提供网络服务的重要组成部分，对5G协议进行安全检测和脆弱性分析具有重要意义[1]。

第3代合作伙伴计划（3GPP）安全保证规范（SCAS）[2]、通信监管部门和运营商均有5G新空口（NR）协议安全测试要求。5G NR协议安全漏洞挖掘是目前的研究

基金项目：国家自然科学基金项目（U23B2003）；广东省重点领域研发计划项目（2020B0101120003）

热点[3]。然而，由于网元数量庞大、设备源代码无法获取等，传统的白盒测试和代码审计方法在5G NR协议的安全检测中已经失效。其次，现有的5G NR协议测试是针对功能和性能的测试，而针对那些更深更广层次安全问题的技术方案仍比较缺乏[4]。此外，在5G NR协议测试中大多聚焦应用层L3协议，针对数据链路层L2协议的研究不多，测试的完整性不高[5]。因此，实现5G NR三层协议的自动化安全测试是一个急需解决的难题。

模糊测试是一种通过向目标系统提供非预期的输入并监视异常结果来发现软件漏洞的方法，能够在不了解相关源代码和逻辑流程的情况下进行黑盒测试，因此被广泛应用于网络协议和可执行文件的安全检测[6]。现有的通信网络协议模糊测试方法主要基于对3GPP技术规范的手动分析，耗时长且资源消耗高。在变异策略阶段中使用简易的变异策略来生成测试用例，例如位翻转和字节算术，无法根据数据情况动态调整变异策略，这导致测试用例的有效性较低。

目前5G NR协议模糊测试实现了无线资源控制（RRC）和非接入层（NAS）的L3协议的漏洞挖掘工作，而在MAC、RLC和PDCP的L2协议模糊测试中进展缓慢，主要存在以下问题与难点：测试需要高效算法来控制数据链路层的通信，涉及任意修改数据包字段，这对算法提出了较高的要求，需要针对数据链路层协议及数据格式等特点设计策略；在复杂的5G协议通信中，检测基站的无效或不符合规范的响应需要全面的模糊测试和验证策略；测试需要利用上下文信息，如安全配置，这只有在实时通信中才能获得，需要结合5G终端设备；需要优化模糊测试算法，以提高协议状态覆盖率和测试效率。此外，模糊测试在5G NR协议测试中遇到最大的问题是时延限制，这是因为5G数据链路层的测试需要满足低延迟要求，以确保实时通信。拦截和转发数据包的时间需要控制在有限的传输时隙内，无法满足5G业务时延性能要求将导致测试用例无法正常测试，因此需要设计特定的协议报文解析和报文处理方式。

为了解决上述问题，我们提出了一种针对5G NR L2协议安全的模糊测试系统。该系统可以根据5G NR L2协议特征设计针对性的变异策略，满足5G业务时延性能要求，提高测试用例的有效性，高效挖掘5G NR协议的安全漏洞，从而提升5G基站的健壮性和安全性。为了评估系统方案的正确性，我们设计并实现了5G NR L2协议的模糊测试原型系统，并在真实的5G gNodeB上进行了漏洞挖掘和性能评估。本研究中，我们的主要贡献总结如下：

1）提出了一种5G NR协议漏洞的自动检测框架，可以覆盖更底层的L2协议安全，实现MAC、RLC以及PDCP协议的模糊测试；

2）针对5G NR L2协议特征，设计高效的模糊测试数据变异策略，以及多种模糊测试工作模式；

3）基于真实的5G网络设备和5G基站，实现了整套原型系统，包含服务端子系统和移动终端子系统；

4）实验数据表明，我们的数据包操作的处理时间优越，能够满足5G业务时延性能要求。此外，系统能够有效检测出MAC、RLC以及PDCP协议漏洞。

1 5G模糊测试技术

在通信网络中，无线接入网络（RAN）为用户设备（UE）提供无线通信服务。RAN由无线基站组成，使用的无线接入技术称为新空口技术。5G NR协议分为3层，即物理层、数据链路层和网络层。其中，数据链路层是本文的研究重点，对应MAC、RLC和PDCP，主要功能是信道复用和解复用、数据格式的封装、数据包调度等。完成的主要功能是具有个性的业务数据向没有个性的通用数据帧的转换。

随着5G移动通信网络的大规模部署和应用，5G移动通信领域出现了一波安全研究浪潮。在过去几年中，研究人员发现3GPP规范中存在许多设计缺陷[7–8]和协议漏洞[9–11]。如3GPP技术规范33.501中所述，大量预认证消息通过未加密的格式发送，可被用来发起拒绝服务（DoS）攻击，并获取5G中移动用户的位置或其他敏感信息[4]。LIU等在文献[12]中针对5G网络新协议——扩展认证协议–认证和密钥协商算法（EAP–AKA），提出一种基于Lowe分类法的安全性分析模型。HUSSAIN等在文献[13]中通过对5G协议栈进行建模并使用验证工具来发现协议中的设计缺陷，但是这种方法并不针对5G UE实际实现中的漏洞。HU等在文献[14]中通过分析5G核心网下一代应用协议（NGAP），研究其协议格式，提出一种基于分区权重表的选择变异模糊测试算法。WANG等在文献[15]中提出一种面向5G专网鉴权协议——扩展认证协议–传输层安全协议（EAP–TLS）的细粒度形式化建模与验证方案并验证了保密性、认证性、隐私性3类安全属性。POTNURU等在文献[16]中提出了一种针对RRC和NAS协议的模糊测试工具，生成包含所有可能标识符的模糊测试用例，并发现了srsLTE和openLTE两个开源电信项目中的新漏洞。YANG等在文献[17]中提出了一种结合机器学习算法的RRC协议模糊测试系统，在无需事先了解协议实现的情况下捕获和解释数据包，通过自动生成全面的用例集来检测协议漏洞。HE等在文献[18]中提出了一种基于预定义规则的5G NAS协议智能模糊测试算法，通过对NAS协议分析设计动态变异策略，并在开源仿真环境OAI中验证了所提算法在

覆盖率和测试用例上具有较好的功能。WANG等在文献[19]中实现了基于模糊测试的5G RRC协议漏洞检测模型，发现了UE和gNodeB的若干漏洞，最后给出了几项增强5G安全性的对策。目前大多数研究集中在5G NR L3协议，即RRC和NAS，而针对更底层的L2协议如MAC、RLC和PDCP的安全测试深度不够且较片面。主要原因在于，更底层协议的安全检测对模糊测试系统和算法的设计提出了较高的标准要求，需要满足5G业务更严苛的性能需求。为了解决上述问题，我们提出一种适配5G NR L2协议安全的自动检测框架，针对协议特征设计高效的变异策略和测试用例，能够更好地对NR协议进行模糊测试，并在真实的5G网络设备和5G基站中实现了整套原型系统，通过实验测试和性能评估验证本文所提方案的有效性。

2 方案设计

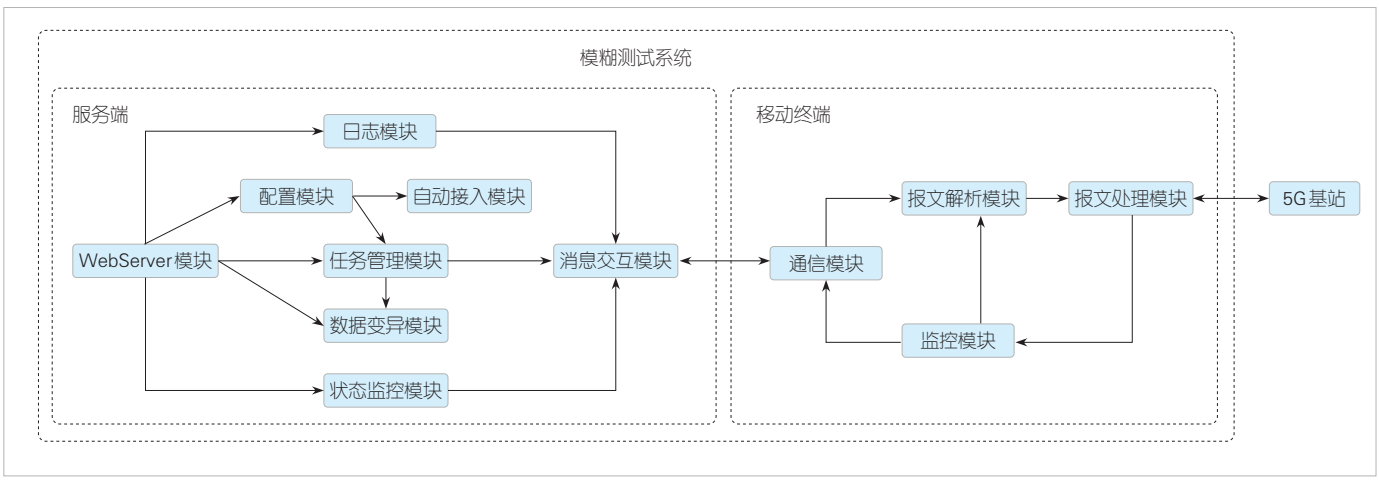
2.1 方案概述

本文提出的基于模糊测试的5G NR协议漏洞挖掘系统架构如图1所示，包括服务端和移动终端。服务端运行在高性能计算机上并提供模糊测试的主要功能，包括数据变异模块、任务管理模块、配置模块、状态监控模块、日志模块、自动接入模块、消息交互模块、Webserver模块，主要提供5G NR协议配置树生成和管理、5G NR协议数据变异、测试任务管理和下发、状态接收处理等功能。其中，数据变异模块和任务管理模块中涉及本文所提方案的关键技术，将在后文详细描述，而其他模块功能只是配合系统整体实现，这里不做赘述。移动终端基于5G通信设备定制化开发功能模块辅助模糊测试，包括报文通信模块、解析模块、报文处理模块、监控模块，主要提供状态监控、共享内存管理、变异数据转发、报文字段解析及变异数据配置等功能。其中，报文解析模块和报文处理模块中涉及本文所提方案的关键技术，也将在后文中详细描述，而其他模块功能只是配合系统整体实现，将不做过多描述。

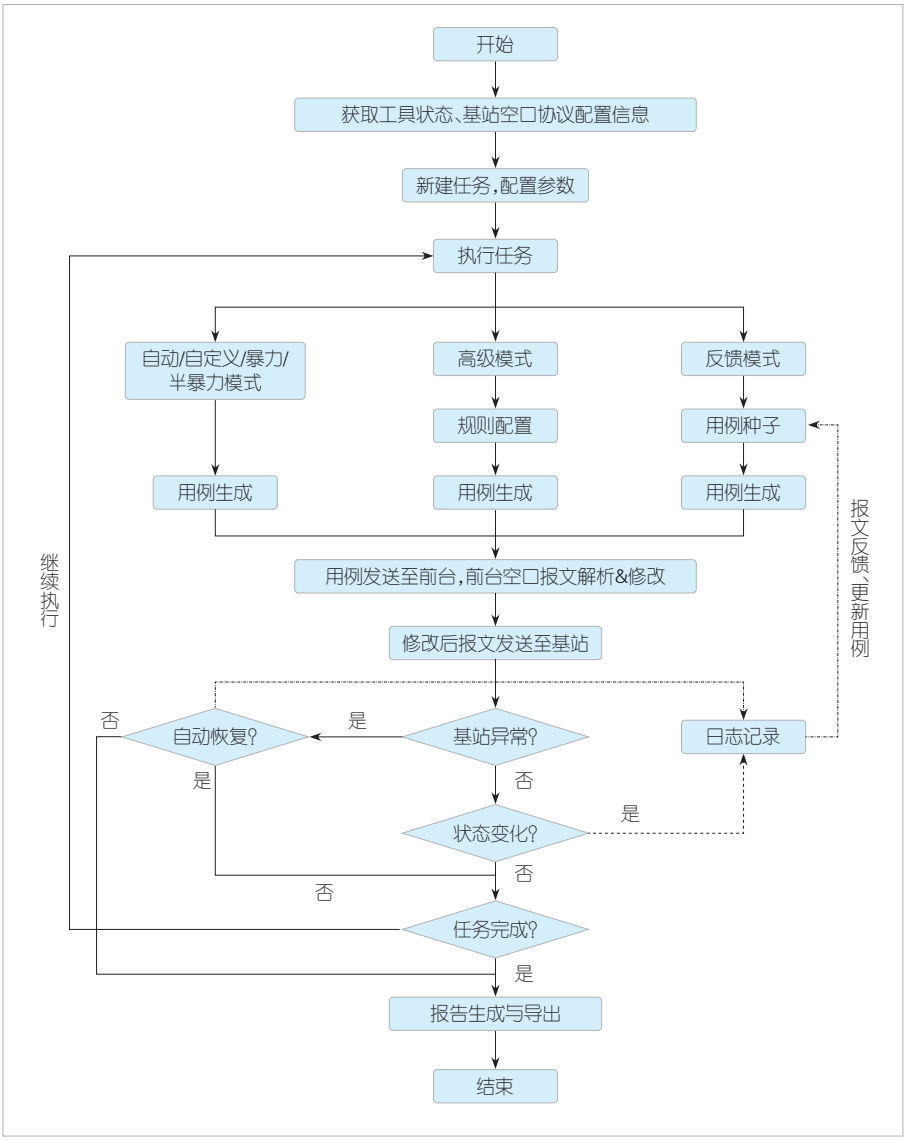
系统开始运行后，首先从5G基站网管系统获取基站空口协议的配置信息。服务端根据基站空口协议配置的协议字段生成5G空口协议配置树（含协议字段名称、类型、长度等），并使用智能变异器对5G空口协议数据进行变异，然后使用服务端转发模块将变异后的配置树数据发送给移动终端。移动终端通信模块接收到变异后的配置树数据后写入共享内存，其中变异数据存储在设备的内存中，报文解析和处理在上层进行，因此需要采用共享内存进行通信。根据获取到的5G基站配置信息，用户在服务端人机交互界面配置将要模糊测试的协议字段。服务端将模糊测试的协议字段配置发送至移动终端。移动终端根据用户配置的协议字段匹配适合的变异配置树数据。移动终端报文解析模块解析MAC层出口的协议数据流，确认协议字段在报文中的确切位置。报文处理模块由共享内存读取变异数据，将变异数据写入对应协议字段位置的数据流，再由物理层处理后经移动终端射频天线发送至5G基站。5G基站收到报文后作出响应。移动终端的监控模块会实时监控5G基站的响应状态，若5G基站业务流程异常，则返回异常状态。监控模块会修改共享内存中的模糊测试状态标志位，将异常状态和导致异常的数据报文记录到日志中，同时将其回传给服务端任务管理模块存储。模糊测试系统工作流程如图2所示。

2.2 服务端

数据变异模块：负责生成针对性的变异测试数据。服务端分析5G NR L2协议特征和业务特征，结合当前基站的协



▲图1 模糊测试系统框架



▲图2　模糊测试系统工作流程

议配置生成数据模型，然后采用多种算法生成变异数据用于测试。该数据变异算法主要依托于对现有5G NR L2协议各字段特性的研究，同时考虑自动化变异算法的效率和可行性的动态平衡。

我们首先分析5G NR L2协议（PDCP/RLC/MAC）具有的特征。5G NR特征主要包括协议特征和业务特征。协议特征如MAC层协议有MAC控制元素（MAC CE）和填充（PADING）两种类型。RLC层协议具有透明模式（TM）、非确认模式（UM）和确认模式（AM）3种不同的工作模式。PDCP层协议有信令无线承载（SRB）和数据无线承载（DRB）两种类型。业务特征则是各协议层在业务上所发挥的作用，例如小区搜索、系统消息、寻呼、测量、随机接入等。接着根据基站的5G NR协议配置选择对应的业务场景数据模型，结合协议特征和业务特征生成应对不同业务场景的5G NR协议数据模型，如小区搜索模型、随机接入模型。

在模糊测试的生成变异测试数据阶段，采用多协议字段同时变异、无序变异和反馈变异等方式生成新的测试数据，如图3所示。多协议字段同时变异的方法可以极大提升用例覆盖率。无序变异的方式通过设置变异规则可实现重放、顺序、倒序用例测试。反馈变异的方法通过针对报文进行监控、记录、分析，利用行为学习与反馈算法提升测试用例的有效性。系统采用的数据变异算法能够自动逐步提高测试用例覆盖率和有效性，进而发现深层次的问题。

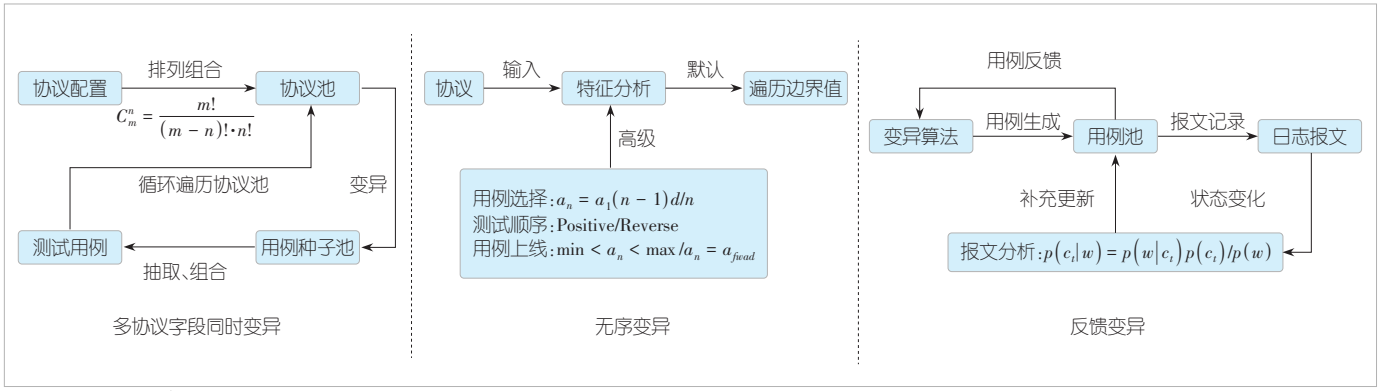
任务管理模块：负责模糊测试任务下发及任务管理功能。测试任务支持自动模式、自定义模式、半暴力模式、暴力模式、反馈模式和高级模式等场景测试。各工作模式特性对比如表1所示。

任务管理模块支持报告查看功能。报告中展示测试的配置及协议组测试结果的详细情况。此外，任务管理还提供模糊测试复测功能，根据需求进行安全漏洞复测，便于开发人员进行漏洞修复。以发现5G基站下行失步漏洞为例，系统发现该漏洞后会存储此时模糊测试使用的业务场景配置、5G NR协议数据模型生成的变异数据以及测试用例。若需要进行漏洞复测，则可以在系统触发任务管理的模糊测试复测功能，系统会使用发现漏洞时存储的信息，按照发现漏洞的流程再次进行模糊测试，进而完成安全漏洞复现。

2.3 移动终端

报文解析模块：负责对5G NR协议报文进行解析。移动终端接入5G基站做数据业务，解析模块按照标准协议报文格式将移动终端5G NR协议组包后的业务报文进行解析，识别出协议层、协议字段及协议字段属性。

系统主要参考3GPP协议格式完成业务报文解析，从移动终端接口获取5G NR协议业务报文，经报文解析后能够识别出协议层如MAC、RLC等，协议字段如逻辑信道标识符（LCID）、序列指示符（SI）、序列号（SN）等，及协议字段



▲图3　数据变异方式

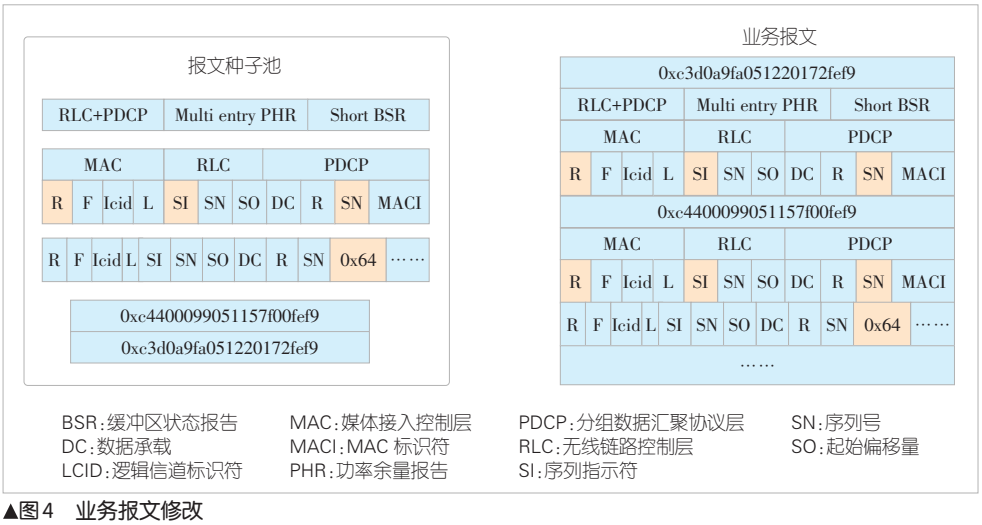
▼表1　各工作模式特性对比

模式	测试粒度	生成方式	特点
自动模式	协议字段	基于突变	持续化，协议字段随机排列组合用例测试
自定义模式	协议字段	基于突变	协议字段指定组合精准测试
半暴力模式	协议组	基于生成	协议组字段变异，粗粒度测试
暴力模式	比特流	基于突变	完全随机化测试
高级模式	协议字段	基于生成	手动针对协议字段设置用例，精细化测试
反馈模式	报文、数据服务单元	基于反馈演进	依据信令变化反馈测试，形成测试闭环

属性如Length等。

报文处理模块：获取根据5G NR协议数据模型自动生成的变异数据，并据此对发送给基站的报文进行修改，经射频天线发给5G基站。变异业务报文数据由报文头和变异数据组成，其中报文头可根据需求定制。

报文处理模块首先从共享内存中获取变异数据，接着根据服务端的模糊测试任务配置完成5G NR协议报文修改。其中，获取和修改5G NR协议报文要满足5G NR协议时延要求。业务调度按照原有的时间窗将修改后的业务报文经物理层及射频天线发送给5G基站。采用字符串匹配算法修改5G NR协议报文，可使获取和修改5G NR协议报文耗时不超出业务调度的时间范围。以反馈变异为例，获取业务报文码流经变异后存储至报文种子池，进而以报文种子池中的报文修改业务报文，完成报文修改。以反馈模式中业务报文一个TB块的若干服务数据单元（SDU）格式修改方法如图4所示。



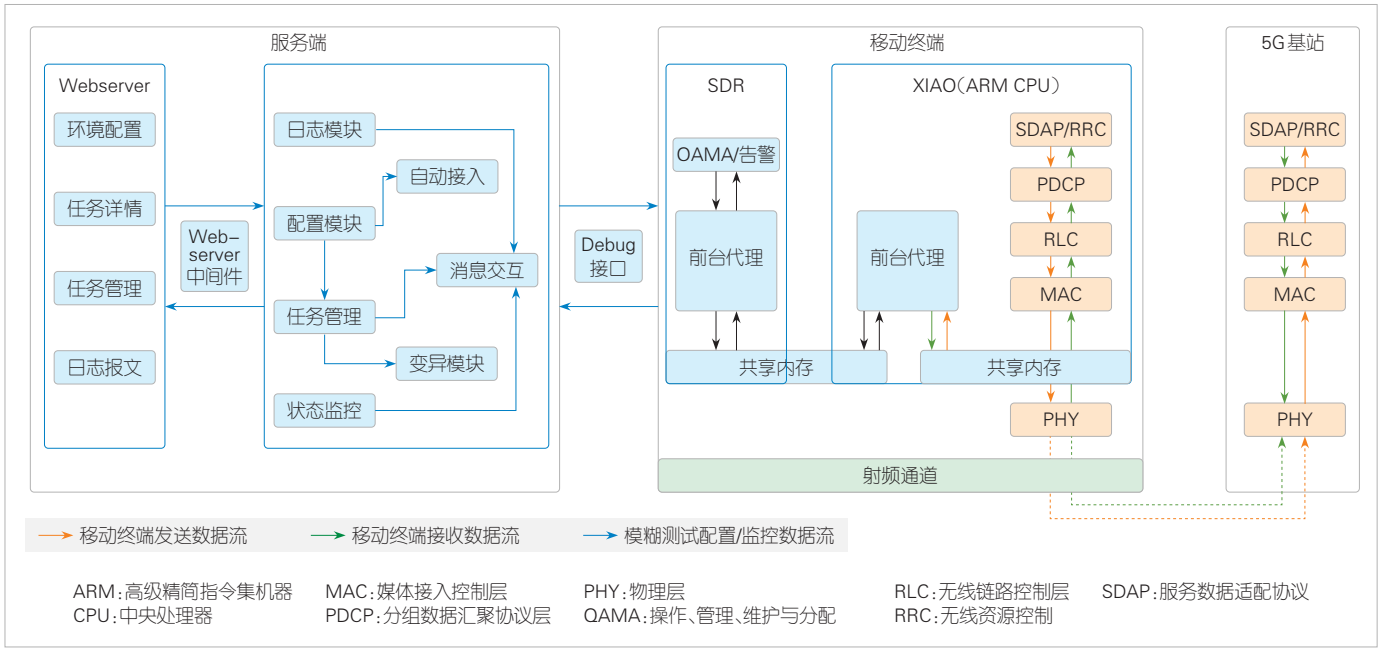
▲图4　业务报文修改

3 实验评估

3.1 实验环境

为了评估本文设计的解决方案，本章节中我们搭建实际的测试平台。测试实验的5G NR协议模糊测试系统架构如图5所示。服务端使用Python语言开发，运行在高性能计算机上，通过网口与移动终端Debug口连接。移动终端的硬件平台基于5G CPE设备深度定制开发，包含两大Linux子系统，分别是SDR系统和XIAO系统。移动终端主要完成SDR、XIAO系统的调度及消息交互。SDR系统主要完成5G NR协议业务流程，包括模糊测试变异数据存储、模糊测试测试控制信息的转发及模糊测试状态的获取及回传等；XIAO系统主要完成5G NR协议的实现，包括NR协议数据解析、NR协议数据处理、模糊测试场景判断及变异数据修改等功能。

5G模糊测试移动终端是基于5G设备开发的，因此在图



▲图5　基于5G设备的模糊测试系统架构

1的移动终端中主要展示模糊测试定制功能模块，移动终端固有支持功能未提及，而图5中移动终端是按照功能模块具体技术实现展示的，包括系统组成、数据流向及通信方式等，其中系统组成包括SDR子系统、XIAO子系统及射频通道。通信模块主要采用核间共享内存通信技术完成模糊测试配置和变异数据传输。监控模块对应操作、管理、维护与分配（OAMA）和告警功能通过该模块获取目标基站的状态信息。报文解析和处理模块均由XIAO系统的前台代理实现。SDR子系统的前台代理完成数据中转及目标状态实时反馈至用户界面。图5中红色箭头代表移动终端发送数据报文到5G基站的数据流向，绿色箭头代表移动终端接收5G基站响应数据报文的数据流向，蓝色箭头代表模糊测试系统配置及监控的数据流向。

测试实验主要工作流程为：服务端根据指定的5G NR协议字段生成协议配置树，并使用变异策略对5G NR协议数据进行变异，然后将协议配置树和变异数据发送给移动终端。5G模糊测试移动终端首先通过SDR子系统接收配置树，并读取变异数据，再将配置树和变异数据写入共享内存。然后，XIAO子系统从共享内存中读取配置树后，解析MAC层出口的协议数据流，读取共享内存的变异数据，将变异数据写入对应协议字段位置的数据流中。数据在PHY层处理后经5G模糊测试终端空口发送至gNodeB。gNodeB收到报文后进行响应。若gNodeB业务流程异常，则返回异常状态。移动终端监控到异常状态后，修改片内共享内存中的模糊测试状态标志位，将异常状态和导致异常的数据报文记录到日志中，同时将其回传给后台监控服务。

3.2 性能评估

我们通过验证模糊测试中解析修改MAC、RLC、PDCP层数据包所消耗的时间来衡量系统性能，并通过详尽枚举协议的模式、类型、SN长度来展示测试的完整性。模糊测试耗时测试分为修改1个TB块的1个SDU和11个SDU数据两种场景。修改1个TB块的若干个SDU数据可根据需求由程序控制。耗时测量结果如表2所示。总体而言，设计方案的5G数据包处理时间远远低于5G业务数据报文一次调度最大时延为200 μs的性能要求。以最复杂的PDCP协议DRB 18 bit SN场景下修改1个TB块的11个SDU为例，测试所使用的时间是4.293 μs，即1个TB块若有1 000个SDU，则在该场景下最多能够修改368个SDU数据，处理能力远高于gNodeB业务需求。

在漏洞挖掘方面，我们首先罗列了MAC、RLC、PDCP的32个代表性协议字段，然后针对性地对每个字段进行了模糊测试，如表3所示。通过我们方案设计的变异策略，如多协议字段同时变异、无序变异和反馈变异等方式，生成畸形的数据包被注入系统，以监测是否发生崩溃，从而进行进一步漏洞挖掘。5G基站上各协议字段的模糊测试结果如表3所示。基站在收到无效的MAC“mac-nr.dlsch.lcid”和RLC“rlc-nr.am.dc”字段后发生了崩溃（gNB Crashed）。此外，由于PDCP字段“pdcp-nr.srb.maci”格式错误，被测试设备

▼表2 NR协议模糊测试测试场景验证结果

协议	模式	类型	SN长度/bit	Fuzz耗时/μs（1个SDU）	Fuzz耗时/μs（11个SDU）
MAC	--	MACCE	--	1.813	4.238
		填充	--	1.634	4.026
RLC	AM	CTRL	无SN	0.407	3.719
		DATA	12	0.509	4.127
			18	0.516	4.266
	UM	分片	6	0.486	4.069
			12	0.506	4.137
		不分片	无SN	0.369	3.649
PDCP	--	SRB	12	0.477	3.817
	--	DRB	12	0.514	4.201
			18	0.514	4.293

AM：确认模式　MAC：媒体接入控制层　SDU：服务数据单元
CTRL：控制信息　MACCE：MAC控制元素　SN：序列号
DATA：用户数据　PDCP：分组数据汇聚协议层　SRB：信令无线承载
DRB：数据无线承载　RLC：无线链路控制层　UM：非确认模式

在一次实例中也发生了崩溃（gNB Crashed）。因此，基站gNodeB容易受到MAC、RLC、PDCP层的DoS攻击。除了崩溃，在测试MAC的“mace-nr. shortBSR. buffersize”和“mace-nr.longBSR.lcg”字段时，用户设备发生了重连（UE Reconnect）和断连（UE cannot Reconnect），从而导致数据传输的延迟增加，以及额外的网络资源和设备资源消耗。此外，设备在重连过程中可能会受到恶意攻击，导致用户数据的泄露或设备被攻击。更严重的是，在测试PDCP“pdcp-nr.drb.reserved”时触发断链（Broken Chain），可能导致数据丢失或损坏，以及生产流程中断或事故等。

4 总结与展望

本文中我们提出了一套主要实现5G NR L2协议的模糊测试解决方案。整套系统主要包含数据变异的服务端子系统和数据处理的移动终端子系统，解决MAC、RLC以及PDCP协议安全漏洞自动化挖掘的难点。针对协议特征采用多协议字段同时变异、无序变异和反馈变异等方式生成高效测试用例，我们定义基于突变、基于生成和基于反馈演进多种工作模式进行5G NR协议安全测试，最后将模糊测试技术融入移动终端设备，结合5G基站设计并实现实验评估系统，验证了本文所设计方案的数据包解析处理性能优越。结果表明，所提方案能够有效解决5G测试数据的时延挑战，有效发现5G NR L2协议中的安全缺陷。与现有技术相比，该方案弥补了5G NR L2协议模糊测试技术的不足，促进了5G NR协议安全漏洞挖掘有效性的进步，提升了5G基站的健壮性和安全性。

未来，我们计划继续改进本文所提模糊测试算法，融入机器学习算法指导数据变异，提高测试用例的有效性。此外，我们将设计包括L3协议即RRC和NAS的安全测试，实现一个覆盖5G NR三层协议的模糊测试系统。

▼表3 5G基站上各协议字段的模糊测试结果

协议	协议字段	是否检测	问题
PDCP	pdcp-nr.srb.reserved	√	
PDCP	pdcp-nr.srb.sn	√	
PDCP	pdcp-nr.srb.direction	√	
PDCP	pdcp-nr.srb.maci	√	gNB Crashed
PDCP	pdcp-nr.drb.reserved	√	Broken Chain
PDCP	pdcp-nr.drb.sn	√	
PDCP	pdcp-nr.drb.direction	√	
PDCP	pdcp-nr.drb.maci	√	
RLC	rlc-nr.am.dc	√	gNB Crashed
RLC	rlc-nr.am.p	√	
RLC	rlc-nr.am.si	√	
RLC	rlc-nr.am.so	√	
RLC	rlc-nr.am.reserved	√	
RLC	rlc-nr.am.dc	√	
RLC	rlc-nr.um.p	√	
RLC	rlc-nr.um.si	√	
RLC	rlc-nr.um.so	√	
RLC	rlc-nr.seqnum.length	×	
RLC	mac-nr.reserved	√	
MAC	mac-nr.dlsch.flag	√	
MAC	mac-nr.dlsch.lcid	√	gNB Crashed
MAC	mac-nr.subheader.sdu-length	×	
MAC	mace-nr.crnti.crnti	√	
MAC	mace-nr.longBSR.buffersize	√	
MAC	mace-nr.longBSR.lcg	√	UE cannot Reconnect
MAC	mace-nr.shortBSR.buffersize	√	UE Reconnect
MAC	mace-nr.shortBSR.lcgid	√	
MAC	mace-nr.single-entry-phr.ph	√	
MAC	mace-nr.single-entry-phr.pcmax	√	
MAC	mace-nr.pre-emptiveBSR.lcg	√	
MAC	mace-nr.pre-emptiveBSR.buffersize	√	
MAC	mace-nr.scell-bfr.ac	√	

MAC：媒体接入控制层　RLC：无线链路控制层
PDCP：分组数据汇聚协议层　UE：用户设备

参考文献

[1] BITSIKAS E, KHANDKER S, SALOUS A, et al. UE security reloaded: developing a 5G standalone user-side security testing framework [C]//Proceedings of the 16th ACM Conference on Security and Privacy in Wireless and Mobile Networks. ACM,

2023: 121–132. DOI: 10.1145/3558482.3590194
[2] 王航, 毛俊, 陈利伟. 5G系统安全测试与自动化 [J]. 信息安全与通信保密, 2023, 21(2): 56–70. DOI: 10.3969/j. issn. 1009–8054.2023.02.006
[3] KHAN J A, CHOWDHURY M M. Security analysis of 5G network [C]//Proceedings of IEEE International Conference on Electro Information Technology (EIT). IEEE, 2021: 1–6. DOI: 10.1109/EIT51626.2021.9491923
[4] PIQUERAS JOVER R, MAROJEVIC V. Security and protocol exploit analysis of the 5G specifications [J]. IEEE access, 2019, 7: 24956–24963. DOI: 10.1109/ACCESS.2019.2899254
[5] SULLIVAN S, BRIGHENTE A, KUMAR S A P, et al. 5G security challenges and solutions: a review by OSI layers [J]. IEEE access, 2021, 9: 116294–116314. DOI: 10.1109/ACCESS.2021.3105396
[6] MAYHEW S R. Fuzz testing architecture used for vulnerability detection in wireless systems [EB/OL]. (2022–05–05) [2024–10–25]. https://vtechworks. lib. vt. edu/server/api/core/bitstreams/3be5d061–041e–48da–81e0–e54c45cde529/content
[7] LANOUE M J, MICHAEL J B, BOLLMANN C A. Spoofed networks: exploitation of GNSS security vulnerability in 4G and 5G mobile networks [C]//Proceedings of International Symposium on Performance Evaluation of Computer and Telecommunication Systems (SPECTS). IEEE, 2021: 1–8
[8] RAMEZANPOUR K, JAGANNATH J, JAGANNATH A. Security and privacy vulnerabilities of 5G/6G and WiFi 6: survey and research directions from a coexistence perspective [J]. Computer networks, 2023, 221: 109515. DOI: 10.1016/j. comnet.2022.109515
[9] MISHRA S. Cyber–security threats and vulnerabilities in 4G/5G network enabled systems [J]. International journal of computational science and engineering, 2022, 25(5): 548–561. DOI: 10.1504/ijcse.2022.126259
[10] HUSSAIN S R, CHOWDHURY O, MEHNAZ S, et al. LTEInspector: a systematic approach for adversarial testing of 4G LTE [C]//Proceedings 2018 Network and Distributed System Security Symposium. Internet Society, 2018. DOI: 10.14722/ndss.2018.23313
[11] NGUYEN V L, LIN P C, CHENG B C, et al. Security and privacy for 6G: a survey on prospective technologies and challenges [J]. IEEE communications surveys & tutorials, 2021, 23(4): 2384–2428. DOI: 10.1109/COMST.2021.3108618
[12] 刘彩霞, 胡鑫鑫, 刘树新, 等. 基于Lowe分类法的5G网络EAP–AKA'协议安全性分析 [J]. 电子与信息学报, 2019, 41(8): 1800–1807. DOI: 10.11999/JEIT190063
[13] HUSSAIN S R, ECHEVERRIA M, KARIM I, et al. 5GReasoner [C]//Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security. ACM, 2019: 669–684. DOI: 10.1145/3319535.3354263
[14] HU Y, YANG W C, CUI B J, et al. Fuzzing method based on selection mutation of partition weight table for 5G core network NGAP protocol [M]//Innovative mobile and internet services in ubiquitous computing. Cham: Springer International Publishing, 2021: 144–155. DOI: 10.1007/978–3–030–79728–7_15
[15] 王跃东, 熊焰, 黄文超, 等. 一种面向5G专网鉴权协议的形式化分析方案 [J]. 信息网络安全, 2021(9): 1–7. DOI: 10.3969/j.issn.1671–1122.2021.09.001
[16] POTNURU S, NAKARMI P K. Berserker: ASN.1–based fuzzing of radio resource control protocol for 4G and 5G [C]//Proceedings of 17th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob). IEEE, 2021: 295–300. DOI: 10.1109/wimob52687.2021.9606317
[17] YANG J D, WANG Y, TRAN T X, et al. 5G RRC protocol and stack vulnerabilities detection via listen–and–learn [C]//Proceedings of IEEE 20th Consumer Communications & Networking Conference (CCNC). IEEE, 2023: 236–241. DOI: 10.1109/CCNC51644.2023.10059624
[18] HE F J, YANG W C, CUI B J, et al. Intelligent fuzzing algorithm for 5G NAS protocol based on predefined rules [C]//Proceedings of International Conference on Computer Communications and Networks (ICCCN). IEEE, 2022: 1–7. DOI: 10.1109/ICCCN54977.2022.9868872
[19] WANG H X, CUI B J, YANG W C, et al. An automated vulnerability detection method for the 5G RRC protocol based on fuzzing [C]//Proceedings of 4th International Conference on Advances in Computer Technology, Information Science and Communications (CTISC). IEEE, 2022: 1–7. DOI: 10.1109/CTISC54888.2022.9849690

作 者 简 介

钟宏，中兴通讯股份有限公司首席安全官、技术专家委员会常委；研究方向为网络安全、数据保护、系统安全、人工智能安全等；曾主持或参与多项国家科技重大专项课题，获多项省部级科技奖励。

夏云浩，中兴通讯股份有限公司网络安全工程师；主要研究领域为安全测评、网络安全等。

张金鑫，中兴通讯股份有限公司网络安全专家，高级工程师；主要研究领域为安全攻防、渗透测试、移动通信安全等。

马致原，中兴通讯股份有限公司资深网络安全工程师；主要研究领域为网络安全、虚拟化安全等。

《中兴通讯技术》第30卷总目次

卷 · 期 · 页

卷首特稿

热点专题

下一代多址接入技术

网络大模型

6G多天线技术

卷 · 期 · 页

6G无线系统技术

网媒融合

卷 · 期 · 页

卫星通信技术

数据通信新技术

专家论坛

企业视界

技术广角

综合信息

《中兴通讯技术》2025年专题计划

期次	专题名称	策划人
1	6G立体覆盖技术	李建东　西安电子科技大学教授 刘俊宇　西安电子科技大学教授
2	智算网络	段晓东　中国移动研究院副院长 李　丹　清华大学教授 虞红芳　电子科技大学教授
3	6G网络安全	刘建伟　北京航空航天大学教授 王景璟　北京航空航天大学教授
4	面向6G的高时效智能机器通信技术	张　平　中国工程院院士、北京邮电大学教授 秦晓琦　北京邮电大学副教授
5	网络中的AI技术	解冲锋　中国电信研究院教授级高工 孟洛明　北京邮电大学教授 崔　勇　清华大学教授
6	新一代光传输技术	陈建平　上海交通大学教授 唐雄燕　中国联通研究院首席科学家

中兴通讯技术杂志社
促进产学研合作青年专家委员会

刊物相关信息

投稿须知　投稿平台　过刊下载　论文索引与引用指南

中兴通讯技术

(ZHONGXING TONGXUN JISHU)

办刊宗旨：

以人为本，荟萃通信技术领域精英
迎接挑战，把握世界通信技术动态
立即行动，求解通信发展疑难课题
励精图治，促进民族信息产业崛起

双月刊　1995 年创刊
第 30 卷　总第 180 期
2024 年 12 月　第 6 期（卷终）

主管：安徽出版集团有限责任公司
主办：时代出版传媒股份有限公司
　　深圳航天广宇工业有限公司
出版：安徽科学技术出版社
编辑、发行：中兴通讯技术杂志社

《中兴通讯技术》编辑部
地址：合肥市金寨路 329 号凯旋大厦 1201 室
邮编：230061
网址：tech.zte.com.cn
投稿平台：tech.zte.com.cn/submission
电子信箱：magazine@zte.com.cn
电话：（0551）65533356

发行方式：自办发行
印刷：合肥添彩包装有限公司
出版日期：2024 年 12 月 25 日
中国标准连续出版物号：ISSN 1009-6868
CN 34-1228/TN
定价：每册 20.00 元