ZTE中兴

面向智算场景的
高性能网络技术白皮书

编制说明

本白皮书在编制过程中得到了多家单位的大力支持，联合编制单位如下（排名不分先后）：

中国联通研究院

中国信息通信研究院

目录

1 前言

智算场景的普惠化正带来新一轮网络技术的革新浪潮。

随着生成式人工智能的发展，AI大模型参数量从GPT-3.5的1.75亿，到GPT-4的1.8万亿，预计未来GPT-5将达到十万亿参数规模，迅速膨胀的AI模型需要更大规模的算力集群执行训练。AI大模型以GPU集群分布式训练为基础，根据阿姆达定律，串行占比决定并行效率上限，网络成为影响算力的重要因素。AI训练任务的高精度并行协同特性以及超大集群互联吞吐量对网络性能提出了数量级的提升需求。AI大模型训练的时间往往长达数月，也使得网络的长稳运行变得前所有未有的重要。从网络流量模型来看，AI大模型训练流量与通算流量呈现出完全不同的特征，突发的稀疏大流成为网络常态，聚合流量具备波峰波谷效应明显、周期性等特征，也使得很多通用数据中心的网络技术不再适用。HPC同样对网络提出高性能需求，特别是在可扩展性以及分布式资源的高效利用方面，HPC与AI的需求趋同。一般来说HPC对于时延更加敏感，但部分采用并行通信的计算模型，同样也关注长尾时延。

AI和HPC集群规模和服务范围的扩大对广域网传输也提出全新需求，包含数据协同和数据快递两大应用场景。数据协同应用主要面向AI/HPC的分布式协同，例如在跨DC的AI训练过程中的是训前模型和数据上载，以及训练期间数据和状态同步过程；数据快递场景包括数据灾备、大规模科学数据传递等。以上都需要广域网具备高性能海量数据传输的能力。

综上，面对大规模AI/HPC的计算、存储和通信需求，不仅数据中心内部的大规模密集数据交换需要高性能网络的支撑，还需要网络能够高效地连接多个数据中心或站点，实现跨地域的AI/HPC业务的高效协同。

本白皮书从面向智算业务的高性能网络需求和技术挑战出发，分析高性能网络技术发展现状和趋势，并探索更适合行业协同发展的高性能网络技术架构和关键技术。

2 术语和缩略语

以下缩略语适用于本白皮书。

缩略语	英文全称	中文含义
AEAD	Authenticated Encryption with Additional Data	带有身份认证的加密算法
AES	Advanced Encryption Standard	高级加密标准
AI	Artificial Intelligence	人工智能

AIGC	Artificial Intelligence Generated Content	生成式人工智能
ARN	Adaptive Routing Notification	自适应路由通知
BBR	Bottleneck Bandwidth and Round-trip propagation time	瓶颈带宽和往返传播时间
BGP	Border Gateway Protocol	边界网关协议
BIER	Bit Indexed Explicit Replication	位索引显式复制
CCO	Collective Communication Offloading	集合通信卸载
CCOM	Collective Communication Offloading Manager	集合通信卸载管理
CNP	Congestion Notification Packet	拥塞通告报文
CPU	Central Processing Unit	中央处理器
CSIG	Congestion Signaling	拥塞信令
DSF	Distributed Scheduled Fabric	分布式全调度网络
DOM	Digital Optical Monitoring	数字光学监控
ECMP	Equal-Cost MultiPath routing	等价多路径路由
ECN	Explicit Congestion Notification	显式拥塞通知
ENCC	End-Network Cooperation Congestion Control	端网协同的拥塞控制
FEC	Forward Error Correction	前向纠错
GCM	Galois/Counter Mode	伽罗瓦/计数器模式
GPU	Graphics Processing Unit	图形处理器
HPC	High Performance Computing	高性能计算
HPCC++	Enhanced High Precision Congestion Control	增强的高精度拥塞控制
HP-WAN	High Performance Wide Area Network	高性能广域网
IFA	Inband Flow Analyzer	带内流分析器
IGLB	Intelligence Global Load Balance	全局负载均衡
IGP	Internal Gateway Protocol	内部网关协议
INT	Inband Network Telemetry	带内网络遥测
IOAM	In situ Operations, Administration, and Maintenance	带内 OAM
JCT	job completion time	任务完成时间
KDF	Key Derivation Function	密钥派生函数
MOD	Mirror On Drop	丢包镜像

MTU	Maximum Transmission Unit	最大传输单元
NCPC	Network-coordinated Path control	网络协同的路径控制
PIM	Protocol Independent Multicast	协议无关组播
RIFT	Routing In Fat Trees	胖树路由协议
RTT	Round-Trip Time	往返时延
ZTP	Zero Touch Provisioning	零接触配置

3 高性能网络关键需求和挑战

3.1 高性能数据中心网络(HP-DCN)

3.1.1 支持超大规模组网是基础

在Scaling Law（扩展定律）的驱动下，万卡GPU训练集群已成为AIGC核心玩家的及格线，智算中心正迅速朝着超万卡级别的规模迅速发展，国内云商如阿里巴巴、百度等陆续宣布具备10万卡集群的支持能力，而Grok-3的训练集群已达到20万卡级别，预计会有更多10万卡+智算集群出现。

如此大的组网规模势必引发网络技术的质变，高性能网络架构的主要功能设计以及性能要求，都需要放到支持超大规模网络的框架下重新考量。

支持超大规模组网的主要挑战包括：

1）交换机接入容量的限制

服务器GPU网卡数量和接口速率在逐渐增加，呈现每两年翻倍的趋势，当前规模商用的GPU服务器网卡接口达到8*400G，支持800G的GPU服务器也已推出，为了满足接入需求，减少设备数量，对单交换机容量提出了越来越高的需求，而单交换芯片的容量提升节奏，明显落后于IO总线的发展，并且存在物理上限。

2）组网拓扑的限制

为了满足数十万卡乃至更大规模的组网需求，在交换机支持端口数短期无法跃升的情况下，传统的CLOS架构需要采用更多的网络层次，更多的网络层次意味着转发跳数增加，在带来更大时延的同时，更多跳数的路径也增大了故障发生的概率和定位难度，使得网络难以

运维；同时，各层级之间用于互联的端口数量剧增，若采用光纤连接，光模块部分的成本增加也不容忽视。

3）异构网络的互通挑战

大规模网络的构建可能会涉及多厂商设备，当前大模型训练网络仍处于技术方案耦合度较高、标准不完备的状态，未来设备间互通可能面临挑战。

3.1.2 超高稳定性是前提

AI和HPC均是典型的分布式系统，网络作为分布式系统的连接底座，网络的故障或者性能波动会影响集群计算效率，因此网络自身的稳定性是整个系统稳定运行的基础。此外大规模训练或计算任务可能持续数周或数月时间，因此要求网络需具备长期持续的稳定性。

高性能网络的稳定性可采用如下两方面的指标衡量：

1）网络可用性：即网络无故障运行的时间，该指标主要与故障间隔时间以及故障恢复时间相关；

2）性能一致性：即在不同网络负荷状态下，网络性能指标具备稳定性与一致性。

高性能网络的稳定性主要面临如下挑战：

1）网络设备故障感知能力不足，协议软件层面的故障感知时延在毫秒级，故障恢复效率低。

2）在大规模网络中，错包和丢包无法避免，链路中如误码产生的随机丢包对于模型训练的影响巨大，但缺乏与高性能组网需求匹配的恢复机制。

由于对带宽的强劲需求，网络需要采用更高的端口速率，但误码率也随之增加，为了纠错而引入的高精度FEC机制又会带来更大的时延。例如，400G及以上速率以太网普遍采用PAM4调制代替NRZ，以在不新增光纤的基础上增加网络带宽，有效提升传输效率；为了应对PAM4带来的比特错误率更高的问题，400G标准中采用了纠错能力更强的RS（544,514）FEC技术，但是这种纠错技术带来了更高的传输时延。基于400G以太网，对比RS（544,514）与轻量级前向纠错RS（272,258）两种FEC技术的时延，前者每包传输的时延约是后者的1.8倍。

针对丢包，现有技术多采用端到端重传机制，对网络吞吐影响较大。已有研究表明，基于RoCEv2的Go-back-N丢包恢复模型，当丢包率达到1%时，RDMA报文吞吐量接近为0。

3）不同负载情况下，网络性能指标波动，性能表现不一致。通用数据中心网络在轻载时一般都可以达到较高的性能，但AI训练是多任务集群，在多任务并存的情况下，同时满足各任务的高性能需求，对于网络资源的规划和保障提出了更高要求。

3.1.3 极致高性能是核心

为了最大化集群算力利用率，AI大模型训练通常采用并行处理机制，将一个任务分布在多个GPU上。并行训练整体分为三个步骤：处理、通知和同步。在处理阶段，每个GPU完成各自的任务部分；在通知和同步阶段，GPU进行卡间通信并汇总出整个任务的结果。整体的JCT由GPU返回计算结果的速度以及网络的同步速度决定。因此，网络性能是影响集合通信效率的重要因素，主要包括以下指标：

1）超低时延：网络时延分为静态时延和动态时延，静态时延为交换机收发报文以及报文在线缆上转发的固有时延，而动态时延由则包括排队时延以及拥塞、丢包引入的时延。在AI大模型训练中，集合通信的网络时延和业务吞吐性能呈现正相关，决定了训练加速比的上限，因此需要网络尽可能降低时延，目标在亚微秒级。

2）极低抖动：网络抖动同样是影响JCT的重要因素，集合通信的流程一般可分解为计算节点之间的多次平行的点对点通信，存在木桶效应，每一步计算中，需要所有节点全部完成点对点通信后，才能进行下一步计算。网络抖动会引发点对点通信时延增加，直接影响集合通信的计算效率。

3）有效高吞吐：大模型训练期间的数据通信体量巨大，可达几百TB级别。机间GPU高速互联，要求网络提供与机内IO总线性能相匹配的单点接入带宽以及网络整体带宽。然而，需要注意的是，对于网络的需求不仅是高带宽，最终目标是业务的有效高吞吐，有效吞吐率需尽可能接近满吞吐。

对于网络极致高性能的需求主要面临如下挑战：

1）时延方面：动态时延一般比静态时延高出几个量级，是网络低时延的主要矛盾，由拥塞、丢包引起的时延往往在毫秒级，因而需要精度更高的流控/拥塞控制及故障恢复机制。

2）抖动方面：现有拥塞控制技术如DCQCN等，主要面向丢包和吞吐，对于控制抖动的考虑较少，需要更精准的拥塞及流控机制。另一方面，针对大模型训练产生的大象流，网络负载均衡粒度过粗也会导致网络抖动难以管控。

3）吞吐方面：高吞吐设计是一个复杂的系统工程，涉及机内外和软硬件的精细协同，单项指标的提升无法有效提高吞吐率，还可能导致此消彼长效应，丢包、时延、吞吐往往会相互影响。

3.1.4 多维自动化运维体系是必需

与通用数据中心网络相比，HP-DCN网络具有节点链路规模大、流量突发性强、集群系统复杂等特性，在网络运维领域，对运维体系提出了以下需求：

1）广度——全面覆盖每个潜在故障点。AI大模型训练网络中的任何链路问题都可能产生广泛影响，鉴于GPU集群网络中链路数量和等价路径数量庞大，传统的Fullmesh监控方式难以覆盖所有链路。未监控链路上的故障会导致故障发现时间显著增加，定位难度也随之加大。

2）精度——捕捉突发拥塞。AI大模型训练产生的突发流更为同步、粒度更小。传统的秒级或亚秒级监控无法满足对链路状态实时测量的需求。例如，在一秒内可能经历多次持续时间仅为毫秒至数十毫秒的吞吐量波动。在秒级尺度下观察时，链路似乎处于稳定低负载状态而无明显波动。这种表象与实际情况存在巨大差异，在逻辑上将导致不同的故障发现和定位结果。

3）深度——节点级别的根因监控实现快速排障和自愈功能。对于承载AI大模型训练流量的ROCE高性能网络而言，一旦检测到异常情况，运营系统必须深入分析以确定根本原因。这包括多个数据源之间的关联分析以及因果关系推理。系统需追踪复杂的关系链并找出最初的触发点，在监控层面直接给出根本原因分析结果，进而加速高性能网络的故障排查和自我修复过程。

尽管通用数据中心网络已采取诸如性能监控、告警系统、电信级OAM等运维策略，并引入随流检测技术以增强运维能力，在面向AI高性能网络运维时仍面临如下挑战：

1）业务视角受限：当前随流检测技术仅停留在网络域，并未延伸至业务端到端层面，从而限制了对业务质量的全面评估和故障定位能力。

2）故障诊断深度不足：在网络故障检测方面存在局限性，尤其是在处理静默丢包和拥塞等复杂问题上缺乏准确判断其根本原因的能力。

3）资源监控缺陷：队列和端口资源监控范围有限且精度不足；对于RoCE网络而言，在满足基础运维需求的同时还需要实现RoCE指标的可视化管理。

4）流量监控精度低：现有流量监控采样机制精度不高，无法实现对流量的精准监控与全面覆盖；这直接影响了网络性能优化效果以及故障定位准确性。

针对AI大模型训练网络的特定需求与挑战，必须建立一个多维分层可观测运维体系：多维度包括实现监控、调优与排障三大核心维度能力；分层覆盖业务层、网络服务层以及基础

网络层。旨在确保从不同层面全面洞察系统状态，为三大核心维度能力提供关键数据支撑，有效应对各种运维挑战。

3.1.5 可规模扩展的安全机制是保障

在计算集群规模日益扩大及网络智能化发展大背景下，网络安全的重要性愈发凸显。将网络安全融入到业务流程，以实现对用户敏感数据保护，传输数据的机密性及完整性保护，以及提升网络安全的可用性，是实现网络安全运营，应对数据泄露及隐私侵犯等风险的必要手段。

随着网络云化的加速，传统的物理控制管理已无法满足现代云环境中的安全需求。在大规模数据中心中可能面临的网络威胁包括：拒绝服务（DoS）、伪造请求源和修改请求消息数据等中间人攻击。

基于大规模智算中心网络的安全管理场景，传统的IPsec和MACsec协议需要通过集中化管理、动态SA管理、实时监控和分布式架构等方法进行优化，以提高网络安全性能和管理效率。这将有助于确保在大规模部署中实现高效、安全的网络通信。在传统安全协议进行优化或调整的同时，应满足以下需求：

1）性能增强：通过优化加解密钥算法的应用，有效降低内存消耗、优化资源使用，并减少对专用硬件的需求，从而降低整体实现成本。

2）安全功能增强：提供增强的加密和身份验证功能，确保加/解密密钥的安全性和不重复使用，增加数据机密性及完整性保护，提供网络安全可用性。

3）支持规模扩展：提供安全密钥生成及维护可扩展性支持，降低安全处理对硬件资源的需求，解决安全源依赖性，支持大规模安全密钥状态及维护。

在满足以上需求过程中可能面临如下挑战：

1）组网规模带来的资源消耗剧增： 智算中心计算量呈几何级数增长及网络规模的快速演进，要求新安全技术在实现规模支持同时，有效降低对硬件资源依赖；

2）用户多样化需求复杂：智算中心网络用户需求多样化，包括可靠、定序、重传等不同服务类型。为满足差异化服务的安全传输，需结合报文序列号检测，数据包完整性验证及重传检测等技术，通过安全协议与数据链路层及传输层等技术协同实现对特定安全服务支持；

3）用户隐私数据泄露及滥用风险增加：随着AI大模型对数据需求不断增加，用户数据面临更高的泄露及滥用风险。需加强对用户数据保护，确保信息交换中用户身份和敏感信息不被泄露，强化数据安全管理和隐私保护，以确保数据的安全性和可靠性。

3.2 高性能广域网（HP-WAN）

HP-WAN应用场景中的业务特征涵盖大量数据流动态突发、多并发业务协同传输、通过站点或数据中心之间的长距离连接。主要业务需求如下：

1）支持海量数据和大象流传输，总数据量为10Gbps~1Tbps；

2）基于任务的数据传输，频率可变，例如定期可预测性数据传输；

3）在一个或多个站点或DC之间的长距离传输，最长可能超过1000公里；

4）即时传输，需要立即或在特定时间传输；

5）及时传输，有完成时间但没有实时传输要求；

6）降低平均数据传输成本；

7）保证数据安全性和完整性。

为了满足以上业务需求，HP-WAN面临新的需求和挑战。

3.2.1 有效高吞吐量是焦点

广域高通量传输的首要需求是超高的网络有效吞吐量，影响网络吞吐量的主要因素包括丢包和时延，网络需要提供极低的丢包率及微秒级排队时延以支持高吞吐需求。

广域网实现有效高吞吐主要面临如下挑战：

1）传统在端侧实现的基于端侧调速的拥塞控制机制，由于广域网长链路时延和慢反馈回路导致吞吐量下降及资源利用率低。当发生分组丢失或拥塞时，端侧可能无法基于来自网络的慢反馈及时调整发送速率；

2）由于端侧无法预测网络传输能力，端侧调速效率很低，产生锯齿效应，导致收敛时间长。

3.2.2 公平共享高带宽是关键

除了满足单个传输任务的高吞吐需求外，广域网还需要公平共享链路带宽资源，保证整体的网络带宽利用率和数据搬运效率。

主要面临的挑战为：

1）广域网被动执行贪婪传输，没有可预测的资源调度机制，突发大流量可能导致瞬时占用大量网络资源，导致整体带宽利用率低；

2）网络不区分流量传输的数据量，可能引发大小流之间带宽分配不公平，小流的传输时延增加。

4 高性能网络技术架构

4.1 高性能网络技术现状和趋势

技术现状

当前智算中心网络相对通用数据中心网络最显著的架构特征是端网融合，即网络内生架构纳入端侧算力特征要素，经典技术机制如精细化负载均衡、基于量化信用的拥塞控制、基于集约范式的在网计算等。同时，端侧网卡引入兼容网络拓扑的多路径控制、逐包喷洒等机制。端网在架构上的深度融合，双向加强，从根本上大幅提升了面向智算中心网络传输和连接性能，实现了超低时延、超大节点组网规模、超高传输通量。从设备形态上看，网卡、网络转发节点均属特别设计的专用组件和设备。从协议和信令上看，自物理层、链路层直至传输层，均进行了加强型设计甚至独立设计。

广域网方面，由于存在长距离传输导致的硬时延约束，当前跨数据中心的高性能无损传输范围主要在300公里以内，主要依赖于硬件设备和专用网络保障（例如光纤直连）;面对更长距离的传输需求，目前更多通过静态配置和调度实现数据尽力而为的传输。

发展趋势

随着训练集群朝十万卡以上的超大规模演进，智算中心网络架构进入新的演变和迭代周期。智算交换机端口速率将由800G跃入T时代，单维度负载均衡将演进至包含L2、L3、L4的多维度负载均衡，算法无关拥塞控制机制将灵活兼容多种端侧拥塞算法，使能端网更精密协同基础上的组件解耦，支撑更具成本优势和持续性的产业生态。基于硬件快速检测的全维度性能可观测技术，也势必成为支撑下一代智算中心网络高效运行和交付的关键技术。

高性能广域网无损与有损路线并存。广域无损技术能够为业务提供低延迟、低丢包和高带宽利用率的数据传输服务，除光互联方案之外，确定性网络技术也可用于提供广域长距无损承载能力。由于广域无损对网络有极高要求，对于时延不敏感的业务，也可增强网络能力提供广域有损的数据传输服务。基于IP的高性能广域传输方案能够以更低的成本支持更长的传输距离，基于网侧主动拥塞控制和配额协商，进一步增强端网速率协同，具备满足大容量限时传输的广域高性能传输需求的潜力。

无论是HP-DCN还是HP-WAN，在发展初期，由于缺乏产业生态的支撑，端算网封闭方案成为一种阶段性权宜选项。然而，近年随着智算产业的迅猛发展，端、算、网各方以各自技术积累推出系列产品和方案，健康高效的产业生态正在迅速成型。面向行业标准化解决

方案的产业联盟、标准组织如UEC、IETF、ODCC等均在积极投入和研究，部分已经取得积极进展。基于行业通用标准的端、算、网解耦方案，势必使能更高效能、更优成本的高性能网络基础设施。

4.2 中兴高性能网络技术架构

中兴高性能网络技术基于算内、算间和入算三大场景，提供可支撑超大规模、高吞吐、高可靠、低时延、安全、智能和灵活扩展的高性能基础网络，并具备高度可运维能力。网络通过和端侧、算侧深度融合，进一步提升端到端整体性能。

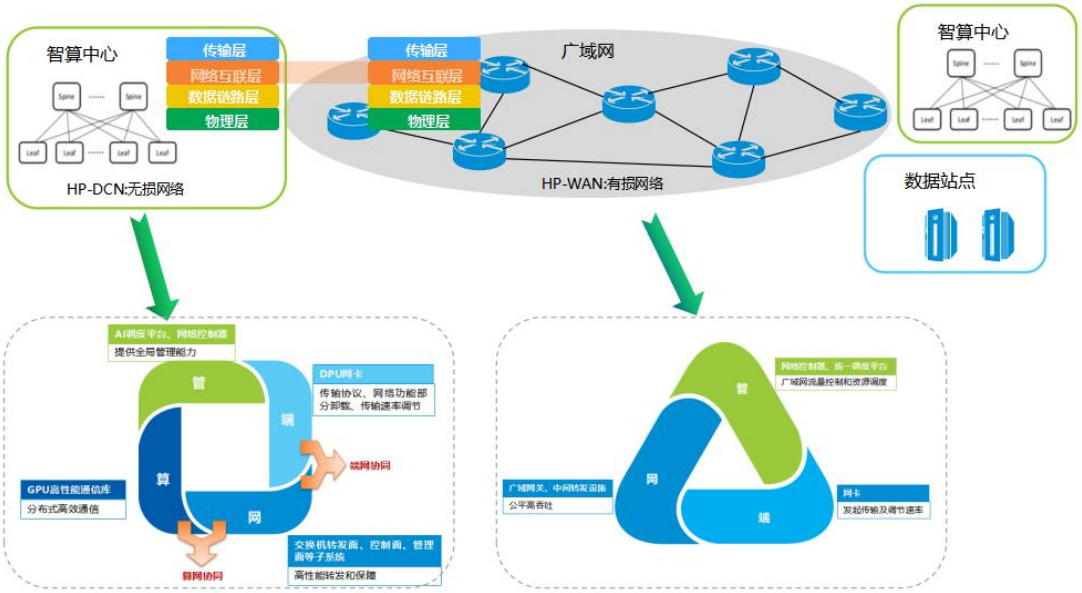

图 1 高性能网络架构整体视图

高性能网络整体架构如图1所示，HP-DCN和HP-WAN之间通过网络层互通。

- HP-DCN为无损网络，包括算、端、网、管四要素，其中：

1）算侧在GPU完成高效计算的同时提供高性能通讯库，实现分布式节点之间的高效通信；

2）端侧包含DPU网卡以及GPU网卡模块，在支持传输协议和提供部分网络功能卸载之外，还进行传输速率调节；

3）网络侧主要包含交换机转发面、控制面、管理面等子系统，需要提供高转发性能以及网络故障的感知和快速保护恢复；

4）管侧包含AI调度平台、网络控制器等管理组件，负责提供全局的管理能力。

以上四个要素在实现各自的高性能特性之外，还需要进行协同以满足高性能需求，典型的协同场景包括：

1）通过算侧和网络侧的协同，可将分布式计算的通信需求信息与网络拓扑和状态结合，实现智算业务流量路径的规划，在网络全局视角做到所有路径的负荷分担和带宽的充分利用；

2）通过算侧和网络侧的协同，还可将算侧的聚合操作卸载到交换机上，从而减少数据在网络中的传输次数，提升有效吞吐率；

3）通过端侧和网络侧的协同，端侧可感知精细的网络状态，进行精准拥塞控制和网络路径控制；

4）通过网络、端侧、算侧和管侧的协同，实现智算网络端到端的业务级管理。

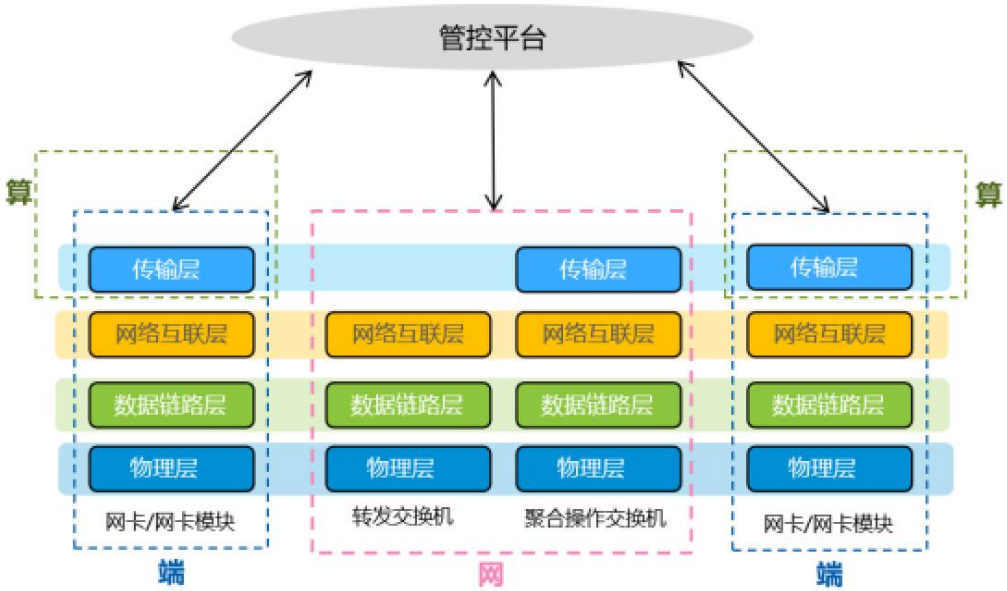

图 2 HP-DCN 网络协议逻辑架构视图

HP-DCN网络的协议逻辑架构如图2所示，各层之间相互支撑，为智算网络提供立体的高性能保障，各层主要功能如下：

物理层：高速率接口、低延时FEC、故障快速检测；

数据链路层：链路级无损、链路级可靠传输、新型转发机制；

网络层：新型负载均衡、新型组网拓扑、路由协议、telemetry技术；

传输层：拥塞/流控机制、丢包重传、网络多路径控制、选择性乱序重排、在网计算、传输层安全。

- HP-WAN为有损网络，主要元素包含端、网、管，其中：

1）端侧：主要是负责进行广域网数据传输的网卡，与DC内类似，端侧作为数据传输的发起端，需要对传输速率进行调节；

2）网侧：包含广域网关和转发设备，在兼顾公平性的情况下提供尽可能高的吞吐；

3）管控：主要指网络控制器和统一调度平台，负责广域网络流量控制和资源调度；

HP-WAN中的协同场景包括：

1）端网协同为主要协同场景，通过协同，网络感知业务流量特征，进行主动拥塞避免，端侧可接收网侧传输建议信息，更快完成初始速率的合理设定；

2）通过控制器和端、网的协同，控制器可基于传输任务特性和网络资源，将资源在不同任务间进行合理分配，提升整体网络的带宽利用率。

在高性能网络的协同流程中，涉及各要素之间的接口和交互，为了支持通用标准化方案部署，各物理单元之间解耦是至关重要的架构设计考虑要素。此外，中兴高性能网络架构也在新型网络拓扑、路由协议优化和安全架构等方向持续探索，为网络高性能提供全方位的技术支撑。

5 高性能数据中心网络关键技术

5.1 超大规模组网关键技术

5.1.1 大规模组网交换机：硬件基础，容量速率双升

传统两层组网容纳GPU卡数量有限，很难满足万卡以上规模组网。超大容量交换机+三层组网模型是支持十万卡及以上规模GPU集群的主流技术路线之一。

51.2T高性能网络为数据中心提供最高带宽密度、最低延迟、最低功耗和最低成本，满足大规模数据中心对于高速、低延迟网络传输的需求，助力构建大规模、高能效的智算网络集群。在实际应用中，51.2T交换机可以通过支持128个400G接口或64个800G接口，来实现51.2T容量的数据交换。

基于51.2T交换机，可采用图3所示的三层clos组网，支持连接132K个400G GPU卡。

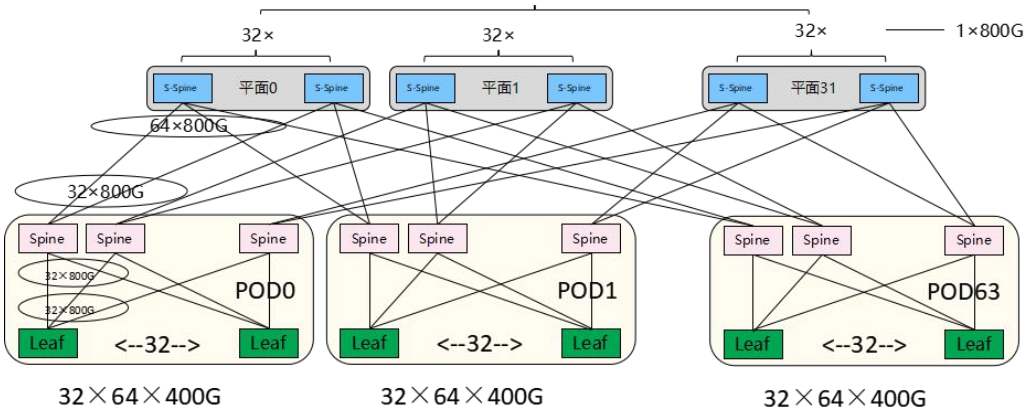

图 3　51.2T 交换机组网示例

800G以太网也成为升级的趋势。当前PCIe Gen5 NIC速度仅支持400GbE链路，而2至3年后的PCIe Gen6 NIC中，800GbE有望用于主流服务器和存储设备。预计800G以太网将进入商用阶段，高速以太网将持续发展。

在图5的组网中，Leaf/Spine交换机可采用64 x 800G交换机，主要搭配使用800G OSFP或QSFP-DD800光模块。根据设备之间传输距离的不同，可采用800G SR8(50m)、800G DR8(100m或500m)或者800G 2xFR4 (2km)光模块来满足需求，灵活满足中大规模AIGC集群运算效能。

5.1.2　大规模组网路由协议：可扩展快速部署，组播能力提供

- 良好扩展能力

拓扑可以支持水平扩展，升级时只需添加更多相同类型的链路和网络设备，无需升级网络本身。

- 协议/部署简单

实现复杂度低，互通性高的协议可以大幅减少网络的维护成本。所选择的路由协议应该通过简单方式就能部署，避免操作维护过于复杂。支持ZTP（Zero Touch Provisioning）的协议更具有优势。

- 故障快速收敛

在网络中有链路或者节点出现故障时，需要支持快速收敛，以便服务的迅速恢复，并且需要严控故障的扩散范围，以免引起整网的拓扑动荡。

- 组播能力

在大模型等智算训练操作中，需要将大规模数据进行GPU间的同步，网络中组播能力的提供，能大量减少相同数据所消耗的带宽，提高数据同步的效率。

5.1.2.1 智算中心拓扑

- **CLOS/Fat-Tree拓扑**

在通用数据中心中，CLOS/Fat-Tree拓扑部署最为广泛，该拓扑中使用Leaf和Spine角色创建无阻塞网络，通过Spine交换机将Leaf交换机（数据中心接入交换机或ToR交换机）互连在一起，将每个Leaf交换机都直接冗余地连接到所有Spine交换机。

两级Spine-Leaf架构中，Leaf交换机之间始终只需要两跳并通过Spine交换机互连。在三级或更多层级的Spine-Leaf架构中，更高一级的Spine交换机将低一级的Spine交换机互联在一起，以提供更高的扩展性。如服务器之间需要更多带宽，只需添加更多Leaf到Spine链路或在架构中添加一个或多个Leaf和Spine设备即可。

由于CLOS拓扑在整个数据中心拥有等距端点，所有服务器之间的通信拥有相同的延迟，该种连接方式也提升了网络通信的稳定性和性能。

CLOS/Fat-Tree拓扑中，Leaf和Spine之间的多条上行链路可通过ECMP方式实现路由转发，避免了对单一链路的依赖而引起的故障。另一方面，如果一个Leaf或Spine发生故障，虽然带宽将减少，但多条可靠的并行链路，可保证服务器之间的通信仍然可以进行。

在智算中心中，CLOS/Fat-Tree拓扑的冗余链路多、带宽无收敛、可扩展性好等特性，使其同样成为应用最广泛的拓扑之一。

- **Dragonfly+拓扑**

Dragonfly拓扑最早在HPC网络中使用，旨在降低网络成本和直径。Dragonfly是一种直连拓扑，其中每个交换机都有一组通向端点的终端连接，以及一组通向其他交换机的拓扑连接，其中一些来自同一组，一些来自其他组，组间拓扑始终是完全连接的。但因组网方式固定，该拓扑存在扩展性弱的问题。

Dragonfly+在Dragonfly的基础上进行了升级扩展，在组内引入类似ClOS/Fat Tree的层次化结构，组间可以采取灵活的连接方式以适配不同的成本和场景需求，可以是组间设备全连接的密集模式和也可以是仅保证组间有连接的稀疏模式。

Dragonfly+拓扑能够适应超大规模的组网需求，具备在AI大模型训练网络中使用的潜力。

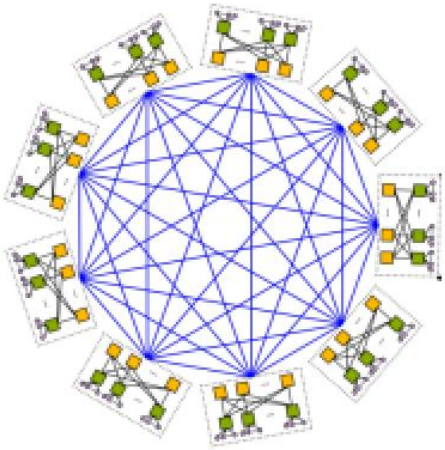

图4　dragonfly+拓扑组网示例

- **2D/3D-Torus拓扑**

环面互连（Torus）可看作是一种网状互连方式，节点排列成 N = 2、3 或更多维的直线阵列，处理器连接到其最近的邻居，阵列相对边缘上的相应处理器连接。典型的Torus拓扑包括2D-Torus和3D-Torus，图5所示：

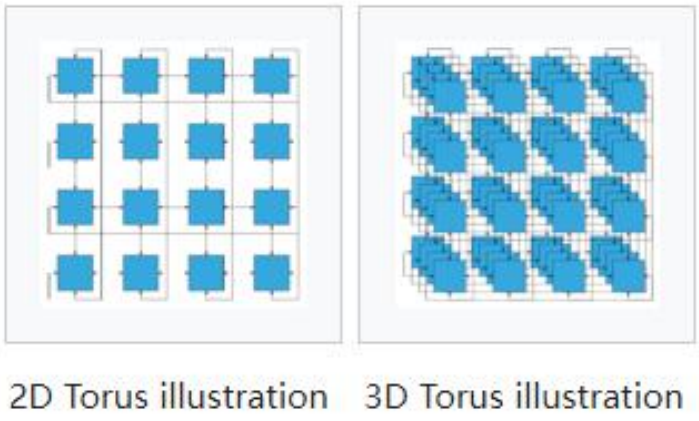

图 5　2D-Torus 和 3D-Torus 组网示例

2D-Torus：节点可被认为排列在正方形的二维格子中，连接方式为，每个节点与其四个最近的邻居相连，同一行或者同一列的两个端节点相连。通信可在二维的四个方向上进行，即+x、−x、+y 和 −y。2D环面的每条边由n个节点组成,总节点数为n2。

3D-Torus：节点可被认为为排列在立方体的三维格子中，连接方式为，每个节点与其六个最近的邻居相连，立方体平行面上正对的节点相连。每条边由 n 个节点组成。通信可

在三维的六个方向上进行，即+x、−x、+y、−y、+z、−z。3D环面的每条边由n个节点组成，总节点数为n3。

2D/3D Torus是最早应用在HPC场景的拓扑形式，国外已有厂商提出将环面互联应用到智算中心拓扑中。

- **Rail-Only拓扑**

Rail-Only拓扑是在AI大模型训练场景中，针对特定计算模式专门进行优化的拓扑连接方式，利用该拓扑能更有效的提升网络性能。

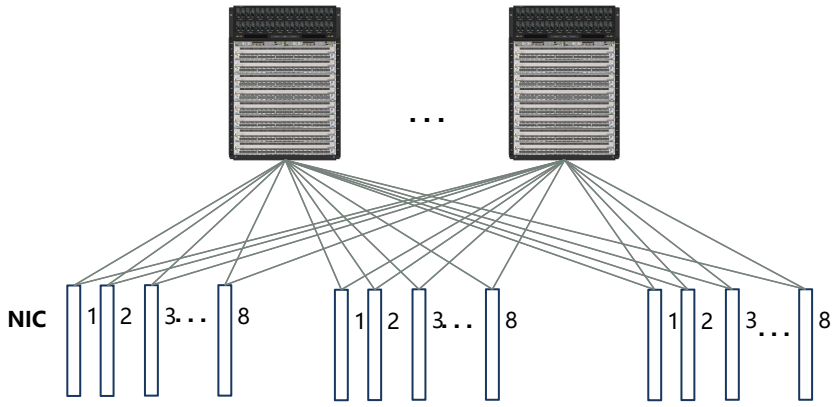

图6 Rail-Only拓扑组网示意

基于LLM训练通信模型具有稀疏性特点，Rail-Only将每组内相同序号的GPU通过特定网络进行连接，能以更低的成本满足GPU间的互联需求。相对于传统的数据中心连接模式，可明显降低网络成本和设备功耗，并能带来更短的时延。

5.1.2.2 智算中心路由协议

- **BGP**

BGP协议是大规模通用数据中心中应用最广泛的路由协议，目前同样可以很好的服务于超过十万台服务器的大型智算中心。

BGP基于TCP协议实现邻居建立和路由交互，其状态机较为简单，路由更新仅发生在路由改变时，仅通告最佳路径，没有周期性的路由数据库同步压力。BGP协议还支持丰富的配置与管理策略，能够根据网络部署的需要进行策略设置，从而达到更佳的控制效果。

- **IGP**

IGP协议包括OSPF和IS-IS协议，是历史悠久的自治系统内部路由协议，但面对AI大模型训练网络向超大规模演进的趋势，IGP面临以下几方面的挑战：

1）规模性：IGP有周期性泛洪的特性，网络规模过大的场景下，泛洪消耗的时间及对链路的影响都将大幅增长，在CLOS/Fat Tree架构下，Leaf及Spine间具有大量平行链路，还会造成大量的重复泛洪，消耗网络带宽。

2）Leaf节点复杂度：IGP路由协议要求整域具有相同的用于路由计算的链路状态数据库，对Leaf节点而言也需要接收和存储大规模数据库用于路由计算，但其实在Spine-Leaf拓扑网络中，Leaf节点不需要整网拓扑状态。

3）策略控制：IGP协议不像BGP协议具有灵活的策略控制功能，因此很难应用在需要策略控制的场景。

4）多平面：数据中心、尤其是智算中心里，采用多平面的部署方式可以充分利用网络拓扑特性，提供更高的路径选择灵活性，但对于传统IGP协议而言则很难部署。

5）新型拓扑适配性：在Dragonfly+等新型拓扑的冲击下，传统IGP协议的优势更难体现。

- **RIFT**

RIFT协议是一种新型动态路由协议，专门针对Spine-Leaf架构设计，能很好的适配大规模智算中心网络。

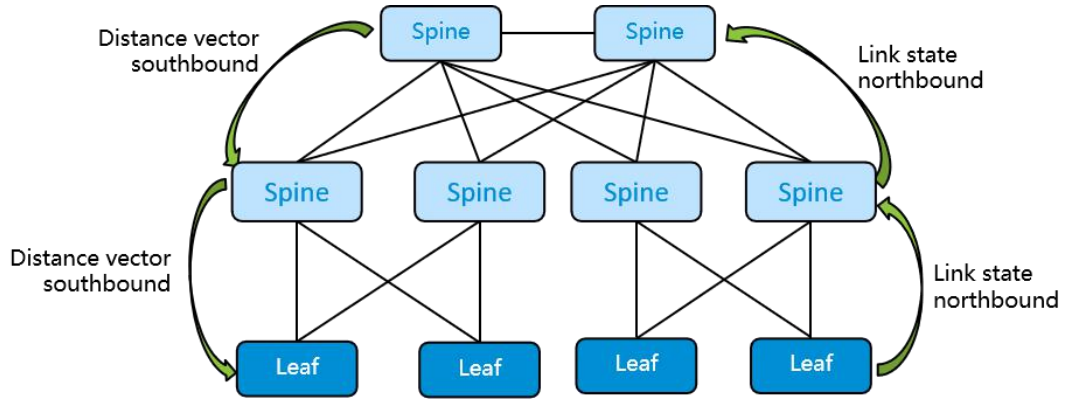

图 7　RIFT 南北向通告

RIFT协议将链路状态协议和距离矢量协议的优点结合起来，具备以下特性：

1）零配置部署ZTP：能够支持设备自动接入网络，并可自动识别可能出现的越级连接错误，大幅降低配置出错的概率。

2）南北向不同通告机制：对于北向节点，则采用IGP类似的链路通告机制，使最北端的Spine节点拥有最全面的路由表项。而对于南向节点，采用类似距离矢量协议通告机制，尤其对于Leaf节点，可仅通告缺省路由，从而使南向节点的状态最小化，尤其大幅降低了Leaf

节点的设备的要求。另外指定Spine节点通告功能，可大量减少北向重复路由的通告，降低对带宽的消耗。

3）解聚合功能：在链路出现故障时，可自动触发解聚合功能，避免可能出现的环路和路由黑洞。解聚合包括正向解聚合和负向解聚合，根据不同的故障情况自动触发。

此外，RIFT协议还具备多链路负载均衡、多平面和灵活策略流控等功能，对CLOS/Fat Tree以及Dragonfly+等网络拓扑具有优良的适配性，因此RIFT协议具备在智算中心网络中使用的前景。

- **BIER**

BIER是一种新型的组播数据转发技术。通过首次将组播业务与组播转发解耦，BIER使组播真正达到层次化的架构划分，网络中无组播状态，网络中间节点无需感知组播业务，仅需根据网络拓扑进行转发，组播业务可以无限增长。对比传统组播协议如PIM等将组播业务流量与转发紧耦合，BIER技术更能应对灵活高效的组播需求。

BIER将组播业务及组播传输彻底解耦，真正实现组播技术的层次化架构，Underlay层面利用IGP/BGP/RIFT等协议进行组播转发层面的构建，同时继承IGP的FRR（Fast Reroute）和LFA（Loop-Free Alternate）功能，在网络出现故障时，收敛速度与Underlay层协议相同，可达毫秒级；组播传输Transport层则是BIER的特色转发层，其创新的数据面能够实现根据拓扑的组播转发，无需单独的信令扩展开销；组播业务仅在网络的边缘设备上由组播Overlay层进行识别，Overlay层可由各动态协议实现，也可由SDN控制器实现，与组播传输完全无关；因此可满足业务无限增长的需求。

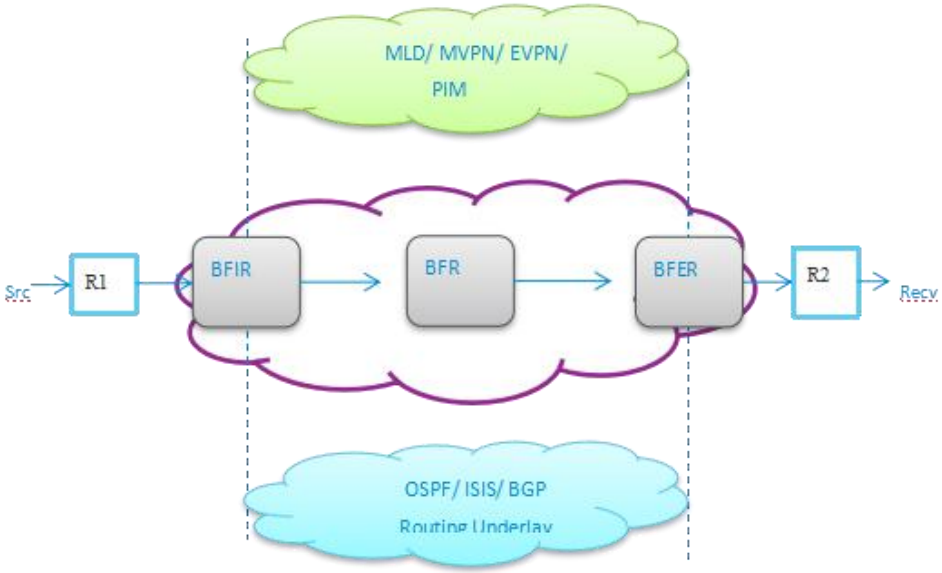

图8 BIER协议架构

BIER技术的业务与转发分离架构，可支持智算中心各类新型业务多点传送需求，流量随到随通，避免耗时的组播树建立及修改过程，能充分适配当前及未来的的智算中心组播业务需求。

5.2 超高稳定性关键技术

5.2.1 故障无感恢复：硬件检测，多级保障

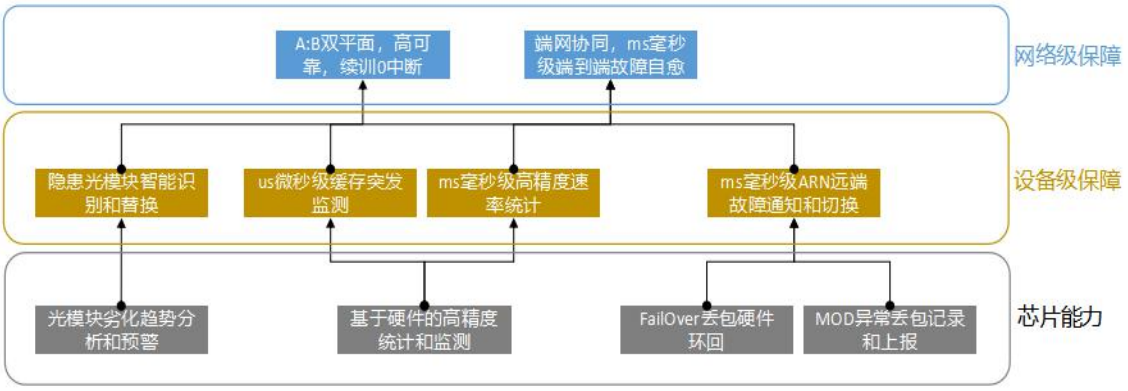

图 9 故障无感恢复功能

如图9所示，为了实现设备级故障和网络级故障无感恢复，需要以芯片能力为基础提供保障。

设备级保障

- 隐患光模块劣化趋势分析和预警

一个智算中心可包含数万块高速光模块，任何一个光模块的故障都可能导致整个AI训练任务被中断。智算交换机基于DOM功能实时监控上报光模块运行状态数据，例如输入和输出功率、温度、电压等；管控平台内置光模块劣化趋势智能分析组件，实时监测各光模块的DOM信息，自动识别光模块劣化风险，在光模块发生信号丢失等故障前提前预警，避免模块故障导致业务受损。

- 微秒级缓存突发监测和毫秒级高精度速率统计

AI大模型训练流量具有高吞吐、微突发、高频次的明显特征，原有基于CPU软件的传统网络性能质量监测技术已无法适用。智算交换机依托自研芯片可编程能力，在芯片各级Pipeline中按需灵活挂接高精度Counter监测指示器，提供端口、队列、流量等不同粒度毫

秒级流量速率统计和微秒级缓存微突发事件监测能力，在整个大模型训练过程中，实时监测网络的性能质量。

- 亚毫秒级故障通知和切换

交换机芯片提供如下能力：

1）硬件丢包环回。对发送给MAC的业务报文，当硬件识别出该端口故障，则会自动将该业务报文环回到发送端，芯片重新查找转发表项获取备用端口发送出去，从而避免丢包。

2）MOD丢包捕捉和记录上报。MOD通过实时捕捉并分析芯片层面各类常见丢包事件（如路由未命中、MTU错误等），精准记录丢包原因及被丢弃报文的关键特征并自动推送给控制分析器。

基于以上能力，交换机的自适应路由功能满足了亚毫秒级故障链路切换的需求。交换机芯片实时监测所有端口/队列转发质量情况，监测到故障或丢包后，首先在本地尝试进行换路，若本地无可用冗余路径，芯片自动生成ARN消息通知上游节点，消息中携带受故障影响的业务报文特征信息；上游节点收到ARN消息时，利用报文特征信息查询本地表项尝试进行换路，整个端到端ARN处理和换路可在毫秒内完成，满足达到亚毫秒级的故障链路切换需求。

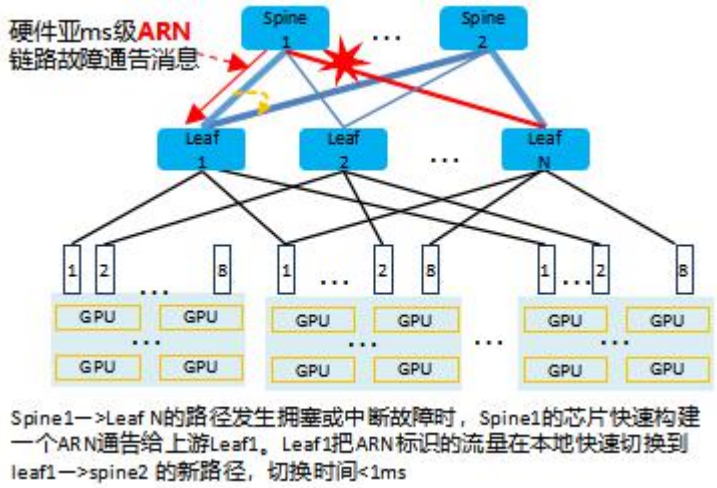

图 10　自适应路由工作流程

网络级保障

- A:B双平面设计

A:B双平面网络架构如图11所示，每张DPU网卡支持2×200Gbps超高带宽，在一个Group中实现GPU数量和通信带宽倍增。双平面设计不仅提升了智算中心组网规模，并且

缓解网卡、光模块、光纤、交换机端口等硬件问题引发的异常，当上行链路或对应交换机故障，流量将无缝切换至另一端口提供服务，训练任务不会中断，仅轻微影响训练速度。

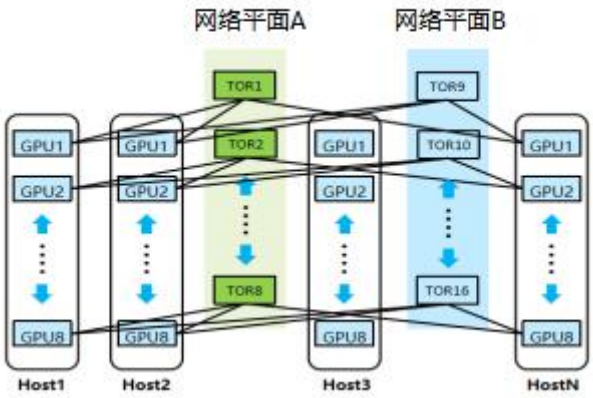

图 11 A:B 双平面网络架构

- 端网协同毫秒级故障自愈

在A:B双平面的网络中，还可以在交换机和DPU网卡上同时开启ARN自适应路由功能，通过端+网协同工作来提升跨平面故障切换性能，实现整网毫秒级端到端故障自愈。

5.2.2 链路级可靠：轻量级 FEC，链路层重传

AI大模型训练网络需要更高的传输速率。对于给定的波特率，四电平脉冲幅度调制(PAM4)能有效地将比特率提高到不归零（NRZ）的两倍，从而提高了高速光传输的效率，并显著降低了PAM4信令传输信道中的信号损耗，因此400G及以上以太网速率采用PAM4编码。

但PAM4信令更容易受到噪声的影响，导致更高的误码率（BER)。假设FEC纠错后的BER为1e-12，则一个包含256个GPU的POD内估计每秒将产生2700个错误帧。虽然PAM4可以采用高级前向纠错（FEC)，以满足更低的BER需求，但更复杂的FEC机制会显著增加延迟。

PCIe和InfiniBand采用了另一种路线，基础思路为，接收器首先使用轻量级FEC（即使用6字节FEC 和额外的8字节CRC保护242字节块）来纠正大部分比特错误，然后检查CRC。如果此检查失败，接收器将启动一个简单的链路层重传协议以再次请求数据。

在以太网上采用低时延FEC叠加链路层重传功能，也能显著降低网络时延，并确保网络可用性。

以IEEE802.3 400GE为例，当前400GE基于PAM4信令并采用RS(544,514)作为FEC方案。FEC纠错前的BER为 2.4e-4，RS(544,514)在物理编码子层的延迟约为62.6纳秒。假设在150米的400 Gb/s以太网光纤链路上传输N个64字节帧，每N帧会有1个帧丢失，1个帧的往返时间（RTT）约为2000纳秒。

但如果将RS(544,514) 替换为RS(272,258)，并对由于采用RS(272,258)的帧丢失率（FLR）较高而导致的额外丢失帧应用链路层重传机制，则能够带来显著的延迟降低。如表1所示，上述场景中，延迟收益计算结果为每1.57e6个帧可以节省4.44e7纳秒。可以看出，链路级重传的延迟成本微不足道，远小于轻量级FEC节省的延迟收益。

表 1: 轻量级 FEC 叠加 LLR 功能的时延收益计算示例

	RS(544, 514)	RS(272, 258) + LLR
码字时延（每帧延迟）	Tcwa = 62.6ns	Tcwb = 34.3ns
FEC 纠错后的帧丢失率 FLR	FLRa =1.7*10^{-12}	FLRb = 6.35*10^{-7}
帧总数	N = 1/(FLRb-FLRa) = 1.57*10^{6}	
重传帧个数	Nra = 0	Nrb =1
N 个帧的延迟代价	Ta = Tcwa*N = 9.83*10^{7}	Tb = Tcwb*(N+1) + RTT*1 = 5.39*10^{7}
N 个帧的延迟收益	Tn = Ta - Tb = 4.44*10^{7} ns	

5.2.3 端网协同的路径控制：端侧传递需求 网络精准控制

网络路径控制（又称为“网络多路径控制”）是由端侧主导的业务路径控制技术，在业务流量性能劣化时，由端侧感知并及时进行路径切换，从而达到保持业务流量性能的一致性、避免拥塞和提高吞吐的目的。

为了实现路径控制的目的，端侧需通过直接或间接的方式感知业务流量路径及路径质量。包括端侧主动探测、端侧模拟计算和端网协同三种技术路线：

- 端侧主动探测路线。由端侧探测路径，进行路径的精确发现和信息维护。通常由发送端主动发送探测报文，通过改变报文 TTL 值以及流量特征值（如源端口号等）实现网络路径的发现，形成发送端的路径数据库。但由于哈希冲突的存在，该种方式有可能导致对网络路径探测不全，此外当出现网络链路变更时，无法及时获取变更后的路径信息。
- 端侧模拟计算路线。基于流量特征值和网络侧流量负载分担算法，由端侧对网络上各设

备选路进行模拟计算，从而得出相应流量的网络路径信息。该方案效率更高，也能够对网络事件更快速的做出反应。但需要端侧提前预知网络设备转发逻辑，并预置算法，端网耦合较紧密，且增加了端侧计算资源的开销和实现的复杂性。

- 端网协同路线。利用网络侧较强的路径探测和控制能力，满足端侧对路径控制的需求。该路线可发挥端和网各自的原生优势，一定程度上对端侧屏蔽网络内部的实现细节，有利于端网解耦，也降低了端侧的复杂度和方案部署成本。

网络协同的路径控制（NCPC）是一种端网协同的实现方式，方案架构如图12所示，主要包括如下元素：

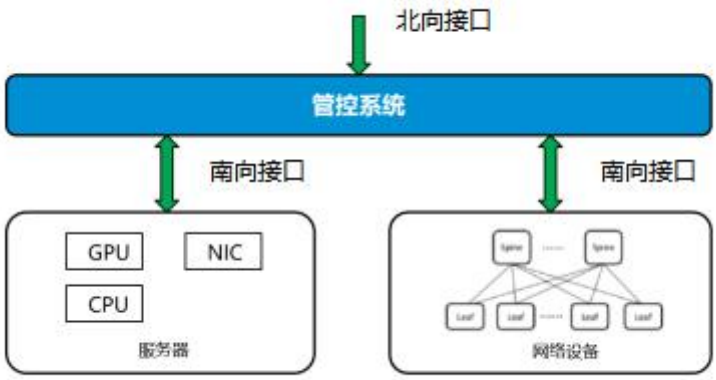

图 12 NCPC 架构

管控系统：纳管服务器和网络设备，感知网络侧服务提供能力。通过北向接口获取算侧业务流量特征和路径控制需求，并为路径需求分配服务标识。通过南向接口基于路径标识在网络侧以路径控制标识为键值进行具体的路径规划和控制策略下发，在服务器侧下发业务流量到路径控制标识的映射关系，以满足相应的路径控制需求。

服务器：网卡根据管控系统配置，在发出业务报文时，携带相应路径控制标识。

网络设备：识别报文中的路径控制标识，将流量引入到相应的路径进行转发。网络侧的内部详细信息，包括负载均衡方式、精确转发路径等在内，均对端侧屏蔽。

NCPC方案中，在网络拓扑、链路、配置等发生改变后，可首先利用设备自身的自适应路由切换、快速故障恢复等机制，对路径进行快速的保障，最终管控系统会基于变更后的网络状态，对各路径控制标识相应的路径进行重新调整和规划，在端侧不感知的前提下，维持网络对外的路径服务能力。

5.2.4 网络隔离与资源保障：网络拓扑隔离，资源合理分配

不同模型训练任务之间需要进行合理的网络资源合理分配且相互不影响，从而保证网络整体的性能一致性。

为了训练任务之间互不影响，通常采用网络逻辑或物理隔离的方式以满足需求。例如IB网络通过Partition Key实现网络隔离，不同租户的IB网络可通过不同的Partition Key来隔离，类似于以太网的VLAN；网络ACL是另一种逻辑隔离的方式；还可以通过多网络平面的方式，在物理上实现网络资源的隔离。此外，需要基于精细化的网络流控机制和QoS策略对网络流量进行优先级划分和带宽限制，确保关键任务的数据传输得到优先保障。

5.3 极致高性能关键技术

5.3.1 层次化负载均衡：整网规划，局部调优，多粒度负载均衡

在大规模复杂网络中，采用层次化的负载均衡方式更加行之有效，通过不同层级的负载均衡的配合，弥补单一负载均衡方案的缺陷，以更好达到全网流量均衡和高吞吐的目的。

层次化负载均衡方案主要包括以下技术内容：

- 全局负载均衡(IGLB)，根据算侧任务流量特征及网络负载状态进行全局路径规划和控制。如图13所示，网络控制器可通过API接口接收算侧调度平台传递的流量特征信息，基于特征信息进行路径的预规划。

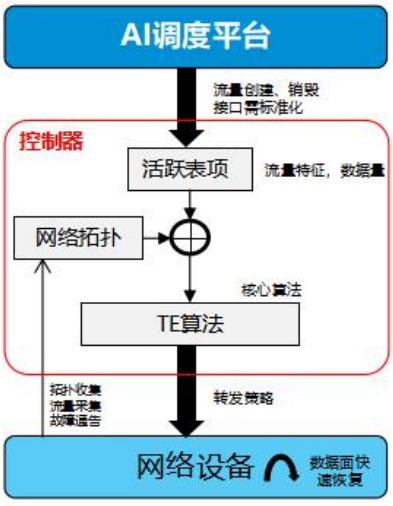

图 13 全局负载均衡流程

- 本地自适应路由，交换机本地根据出口负载状态，动态选择出口。在全局规划的前提下，主要针对网络突发事件引起的瞬时流量不均，及时对路径进行局部调整。

- 远端自适应路由通知，在本地无其他满足条件的可用路径时，通过数据面报文通知上游节点进行切换，完成远端路径快速调整。

  在负载均衡的粒度选择方面，主要的考虑如下：

- 逐包喷洒均匀度最好，但端侧乱序重排需要较大的缓存，同时还要兼顾时延和实现成本；此外逐包喷洒的情况下还需要考虑防止故障半径的扩散，以及不固定的转发路径对网络运维带来的难度。
- 传统的每流基于五元组哈希方式在AI训练场景下容易导致哈希极化和负载不均，但配合层次化负载均衡技术，从全局视角尽可能将大流分担到不同链路上，避免同一时间多个大流共用链路，也可以以较小的实现代价提升网络中流量的均衡度。
- 新型转发技术为网络负载均衡提供了其他的粒度。例如在分布式全调度网络（DSF）技术路线中，流量按报文单元或容器对报文进行转发，通过对报文单元的路径编排，在尽可能保序的同时也提供了较好的负载均衡效果。

5.3.2 拥塞控制：算法无关，迅捷智能

5.3.2.1 智能无损的拥塞控制（AI-ECN）

典型的数据中心组网如图14所示，分布式计算与分布式存储的设计，会导致多个服务器同时向一个服务器传输数据的多打一现象，造成拥塞丢包，严重影响网络的时延和吞吐性能。

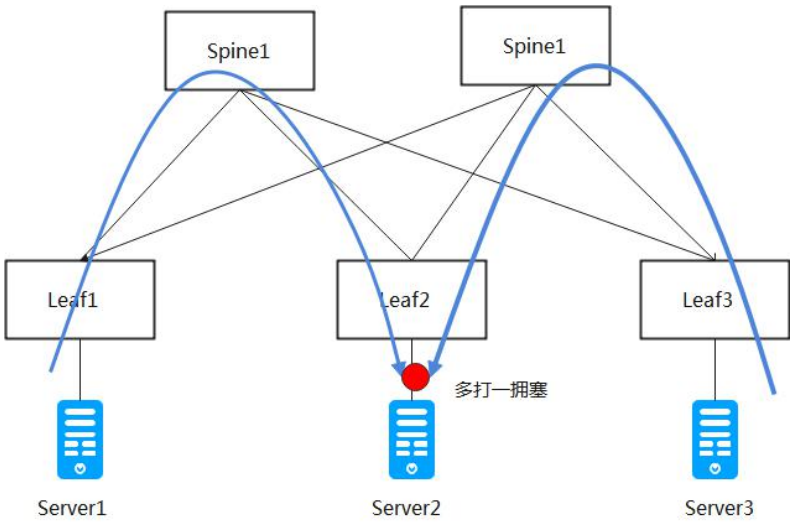

图14　多打一拥塞

数据中心网络通常采用ECN机制进行流量速率控制，流量接收端感知到网络上发生拥塞后，通过协议报文通知流量发送端，使得流量发送端降低报文的发送速率，从早期避免由于拥塞而导致的丢包，实现网络性能的最大化利用。

智能无损功能是对基础ECN功能的扩展和增强，算法模型基于网元智能化架构，利用强化学习与启发式算法，通过实时监控交换机队列的拥塞情况，动态调整ECN水线，实现丢包、吞吐与时延的最优，满足不断增长的大模型算力需求。强化学习算法具有非常好的适应能力，通过离线训练加在线训练的方式，可以应用于多种不同的场景，对于未知的流量模型，当离线训练的算法模型表现不佳时，则启动在线训练，实时学习流量模型，进行精准参数调控。启发式算法通过逐步试探的方式进行参数调整，慢慢逼近最优解，算法计算量极低，所需资源少，具备很强的可部署性，可以轻松支持大规模队列的ECN调控。基于应用场景及可用资源情况，强化学习算法与启发式算法既可以分开部署，也可以统一部署，无论是分开部署还是统一部署，此类智能算法的使用均能带来远优于基础ECN功能的更好的拥塞控制效果。

5.3.2.2 端网协同的拥塞控制（ENCC）

5.3.2.2.1 精细化拥塞控制

无论是基础ECN还是扩展和增强后的ECN，位于IP报头中的ECN字段始终只包含2比特信息，能表示的拥塞信息有限，无法支持更精细化的拥塞控制算法，如业界流行的HPCC++算法。

为了支持HPCC++等更精细化的拥塞控制算法，业界提出了多种带内遥测机制，较为有影响力的包括INT、IFA和IOAM，以及新提出的CSIG等。其中，INT、IFA和IOAM非常相似，属于同一类的长度递增型拥塞控制信号，也即从发送端到接收端长度逐跳增加的拥塞控制信号；携带此类带内遥测扩展头的业务报文在经过每一跳交换机时，都会被添加上诸如队列深度、传输时延这样的拥塞控制信息；这些拥塞控制信息逐跳叠加，直到由最后一跳交换机或接收端反馈给发送端，完成拥塞控制闭环。拥塞信令CSIG则与INT、IFA和IOAM存在较大差别，属于另一类的长度固定型拥塞控制信号，也即从发送端到接收端长度始终保持不变的拥塞控制信号；携带此类带内遥测扩展头的业务报文在经过每一跳交换机时，都不会被添加上新的拥塞控制信息，而是由交换机对扩展头中已有的拥塞控制信息，诸如最小可用带宽、最大节点时延，进行可能的修改；这些已有的拥塞控制信息经过逐跳的可选性修改，由最后一跳交换机或接收端反馈给发送端，完成拥塞控制闭环。上述两类拥塞控制信号的区别如表2所示：

表 2 拥塞控制信号分析

拥塞控制信号格式	拥塞控制信号类型	特点
INT	长度递增型拥塞控制信号	拥塞控制信号的长度随着跳数增加而递增
IFA		
IOAM		
CSIG	长度固定型拥塞控制信号	拥塞控制信号的长度始终保持不变

总体来说，尽管INT、IFA、IOAM、CSIG所携带的拥塞控制信号各异，且分别适用于如BBR、Poseidon、HPCC++等各种各样的拥塞控制算法，当前也各自定义了不同的封装格式，但这些带内遥测机制有着同样的拥塞控制原理。所以采用统一的、标准化的封装格式来兼容各种拥塞控制信号是可行的，也是有必要的。

5.3.2.2.2 快速反馈拥塞信息

随着高性能网络规模的不断扩展，一个高性能网络所覆盖的范围已经不局限于单个数据中心机房，甚至不局限于单个数据中心楼宇或单个数据中心园区。在这些情况下，由流量接收端来向流量发送端反馈拥塞通知已经不能满足极致低时延的要求，这就需要引入快速CNP功能。快速CNP是由检测到拥塞的中间交换机直接向发送端反馈拥塞通告，而无需先把拥塞情况告知接收端，再由接收端向发送端反馈。这种拥塞通告反馈链的缩短在长距离RDMA应用场景下能带来显著的收益。目前，业界已有多种私有的快速CNP技术方案，这些方案一个共同的缺陷就是要求发送端和交换机来自同一供应商，不利于快速CNP方案的灵活部署，所以快速CNP方案的标准化就显得尤为必要。

5.3.2.2.3 流控机制

精细化的流控机制对于控制网络的时延和抖动也至关重要。基于优先级的流控PFC仅支持入口8个队列，单独使用难以满足高性能网络需求，但可以与其他拥塞控制/流控机制配合使用，作为其他机制失效后的最后兜底保障。业界同时也提出了多种新型流控机制，例如端到端的基于信用的流控、链路级基于信用的逐跳流控等。这些流控机制作为高性能网络整体拥塞控制机制的有机组成部分，可与上述拥塞反馈和控制机制配合使用，以达到更好的拥塞控制效果和更优的网络性能。

5.3.3 集合通信卸载：统一编排，轻量传输

集合通信卸载（CCO）技术允许将分布式系统中的集合通信操作卸载到网络交换机上，提供了一种全新的分布式系统通信解决方案。

CCO是是网络领域中用于提升分布式应用性能的技术，能够高效、可控地利用网络设备的存储和计算资源。CCO功能通过将集合操作卸载到CCO交换机上，从而减少数据在网络中的传输次数，降低延迟，提高吞吐量。

CCO框架如图15所示，CCO框架包括以下几个关键组件：集合通信卸载管理（CCOM），基础设施层及CCOM南向接口。

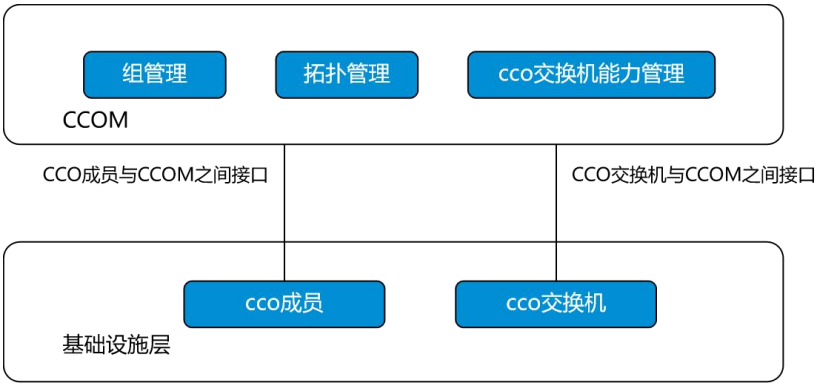

图 15 CCO 框架

CCOM

CCOM负责建立集合通信组，分配必要的CCO资源，并管理CCO交换机资源。CCOM可以通过不同的网络到达CCO交换机和CCO成员。CCOM关键功能模块包括：

- 组管理

  1）创建和拆除集合通信组；

  2）管理集合通信组的状态查询；

  3）分配和释放集合通信组所需的CCO资源。

- 拓扑管理

  1）获取或计算CCO-Tree，CCO-Tree基于底层网络拓扑和CCO交换机的资源信息得到；

  2）管理和维护网络拓扑的变化，以适配集合通信需求。

- CCO交换机能力管理

  1）获取CCO交换机能力信息，包括支持的集体操作类型、支持的组数量、支持的MTU等；

  2）管理CCO交换机资源。

基础设施层

包括CCO交换机和CCO成员。CCO交换机执行集合操作，接收CCO成员的输入数据，执行集合操作，并将输出数据分发给一个或多个CCO成员。

CCOM南向接口

CCOM与CCO成员和CCO交换机之间的交互接口。CCOM南向接口是实现集合通信卸载的关键，它确保了集合操作的协调、资源的合理分配和操作的高效执行。通过这个接口，CCOM能够管理集合通信组的生命周期，从创建、操作到销毁，以及在操作过程中的错误处理和性能监控。

在CCO架构中，基础设施层，CCOM及CCOM南向接口共同构成了一个坚实的框架，使得CCO能够通过CCOM实现统一编排，进而优化集合通信操作。CCOM的集中管理能力在此发挥了关键作用，它负责识别网络设备能力、管理资源，并为集合操作配置网络拓扑等。这种集中化的管理方式极大地简化了分布式系统中的数据聚合流程，提升了效率。

在数据传输的层面，CCO呈现出其独特的优势。CCO采用轻量级传输协议，减少数据在网络中的传输次数，有效降低了延迟。通过在网络交换机中直接处理数据聚合，CCO减轻了端侧负担，提升了数据处理速度，同时减少了因数据传输导致的能耗。此外，CCO的轻量传输还体现在对数据包的高效处理上，它支持头部压缩和精简的控制信息交换，这些优化进一步减少了通信开销。这种设计使得CCO非常适合对延迟敏感的高性能计算和大数据分析应用，它能够在保持数据传输可靠性的同时，提高网络资源的利用率。

5.4 多维自动化运维关键技术：层次化可观测体系，高精度感知

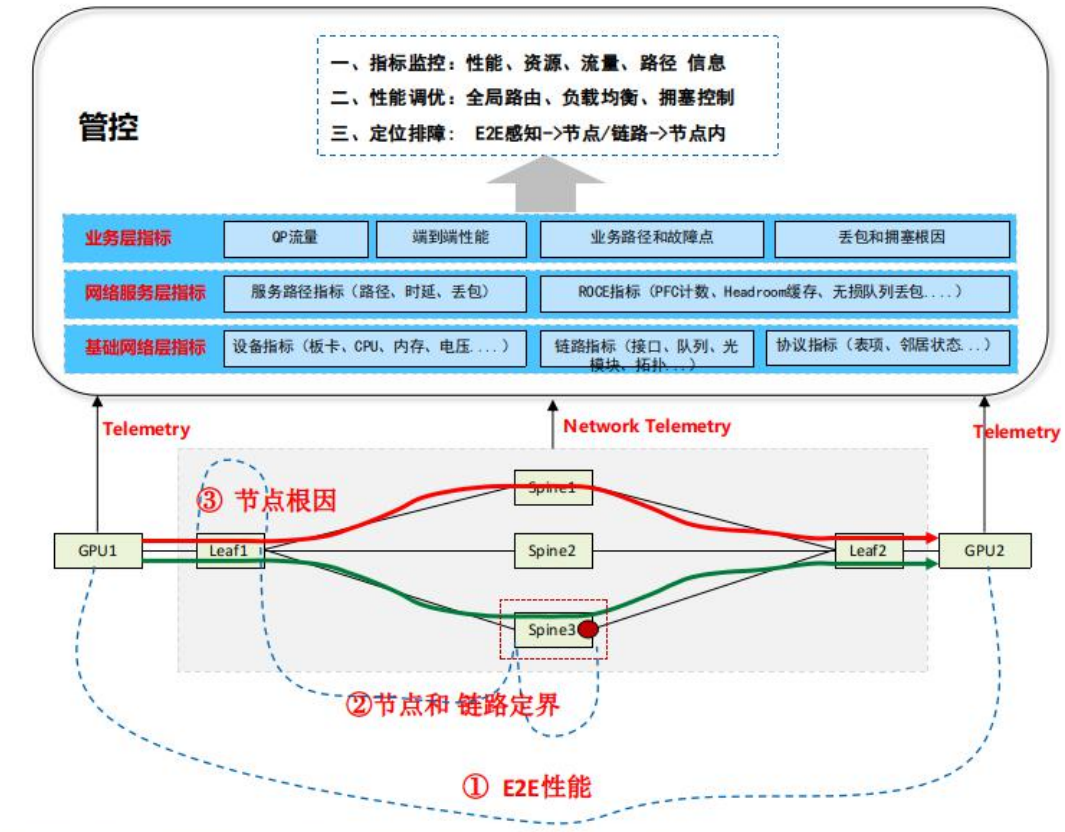

图16　多维分层可观测架构

如图16所示，针对高性能网络的极致需求和挑战，应构建多维分层可观测性体系，并基于 Telemetry 技术，通过算侧和网侧的上报获取网络和业务相关分层指标：

- 基础网络层，主要检测物理网络的健康状态，主要包括三类指标：

  1）设备指标：监控板卡、CPU/内存使用率及电压状态；

  2）链路指标：跟踪接口报文计数、流量统计、队列信息、光模块详情及网络拓扑；

  3）协议指标：检查协议状态、路由及邻居表，确保全面网络监控。通过大数据分析处理上述数据，实现健康评估、风险预测、性能KPI分析及拓扑管理等功能。

- 网络服务层，主要包括两类类指标：

  1）服务路径指标：涵盖ECMP路径的时延与丢包详情，利用电信级OAM机制探测子路径状态，支持业务调整决策;

  2）RoCE指标：监测PFC报文、队列占用情况、Headroom缓存状态及PFC死锁恢复次数。通过大数据分析上述采集数据，实现无损网络故障识别与风险预警，如PFC风暴检测、死锁预防及全网队列一致性检查。

- 业务层指标，通常以业务流为中心，观测业务流性能和路径、故障根因等信息，主要包括以下四类指标：

1）QP流量信息: AI大模型训练中流量模式同步性增强，粒度更细。通过计算或网络节点提供毫秒级业务流统计，包括五元组分析、队列状态监测及逐包字节计数等；

2）端到端性能评估: 获取AI训练场景下业务流实时丢包率与延迟数据，采用随流检测技术（如IOAM、INT或IFA），实现计算侧的全程性能监控；

3）网络路径与故障定界: 记录AI大模型训练业务流传输路径上的所有链路与节点，利用计算侧扩展的随流检测技术实现可视化管理和快速故障定位；

4）根因信息上报: 一方面采用MOD技术，覆盖Ingress、Egress和MMU环节，报告受影响流、事件时间、原因及数量。另一方面采用MOC技术，自动探测微秒级拥塞并精确捕捉突发情况，包括拥塞时间窗口、队列深度峰值及流量特征。

根据上述可观测指标，系统可实现三个综合维度能力：

（一）指标监控可视：实时掌握网络资源与业务动态，如端口占用、队列状态等，通过操作界面直观展示，预警潜在拥塞、丢包等问题，保障资源高效利用;

（二）业务性能调优：监控关键指标，优化网络与业务协同，精准调控拥塞管理及流量调度，确保算力与网络无缝对接，提升处理效能;

（三）故障排查加速：实施多层次故障检测，实现秒级故障定位与分钟级自动恢复，显著增强网络运维效率和可靠性，保证业务连续性。

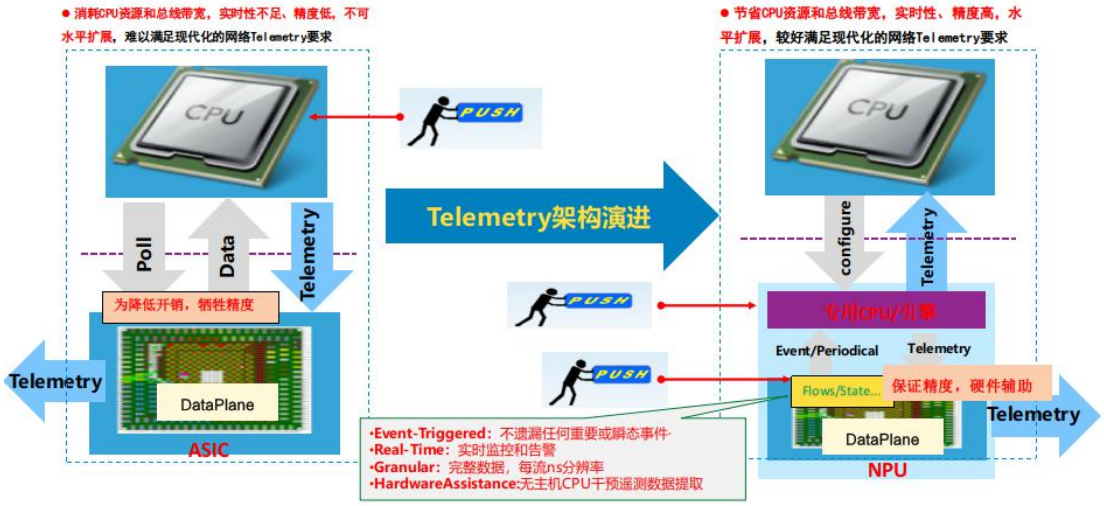

图 17　数据面 Telemetry 技术架构演进

综上所述，智算中心网络的多维分层可观测性需求推动了Telemetry技术的革新。如图17所示，新一代数据面Telemetry需超越传统限制，实现性能飞跃。传统方案通过控制面

CPU周期性轮询采集数据并封装成Telemetry格式，但这种方法效率低下、实时性差且扩展性不足，已无法满足现代网络对高精度、低延迟及可扩展性的严格要求。

面向高性能网络的发展趋势，Telemetry技术将聚焦于以下关键领域：

1）优化芯片算力利用：通过芯片算法与设计创新，在确保信息压缩的同时推动Push源下沉芯片的架构，以实现更灵活的Telemetry格式与封装方式，满足多样化监测需求。平衡带宽消耗、设备资源利用与检测精度是核心挑战；

2）本地感知与智能化决策：全生命周期的业务检测与资源覆盖，在芯片级减少不必要的开销。借助高精度的数据感知结果，驱动本地节点实现智能分析与即时决策机制，显著提升网络效率及响应速度;

3）推进接口与封装标准化：标准化促进设备互操作性与生态融合，显著降低部署成本并提升系统兼容性和稳定性；

4）综合运用多种遥测技术：整合基于流、基于事件及混合型遥测方法，形成全面灵活的解决方案体系，增强网络观测能力，更好地支持智能计算网络的发展趋势。

总之，在构建面向未来的高性能网络可观测体系时，需持续推动芯片内置遥测功能的拓展与智能化升级；同时积极促进行业标准的制定与实施；并不断探索如何将不同遥测技术的优势互补结合，以期为智算网络持续赋能，并最终实现高效、精准且全面的网络运维和管理。

5.5 可规模扩展安全机制关键技术：零信任模型，轻量级加密，安全会话无关

在高性能网络规模快速增长的背景下，通过采用软件定义网络、轻量级安全协议、动态资源管理、容器化架构等策略，可以有效降低对硬件资源的依赖，同时实现新安全技术的规模支持。在安全管理方面，采用零信任安全架构，通过身份验证和访问控制提升网络安全性，保护网络数据及敏感数据不被泄露或破坏。

5.5.1 零信任安全架构

零信任安全架构是一种网络安全模型，其核心理念是“永不信任，始终验证”。在这一架构中，所有用户、设备和应用程序在访问网络资源时都必须经过严格的身份验证和授权，不论其是否在内部网络中。零信任安全架构旨在保护网络数据和资源，防止未授权访问和潜在的内部与外部威胁。

在AI大模型组网中，使用零信任安全架构有助于通过严格的身份验证和授权机制，确保只有经过验证的用户和设备才能访问资源，从而降低安全风险。由于AI模型和训练数据通常

含有敏感信息，通过强制加密和细粒度访问控制，可降低数据在存储和传输过程中的安全风险，防止数据泄露和滥用。

可扩展安全协议遵循端到端的零信任安全架构，即网络上所有设备及容器都不受信任，以规避或减少来自数据中心外部或内部安全威胁。

5.5.2 可扩展安全协议

可扩展安全协议应支持端到端的保密和加密处理，适用于大规模HPC和AI环境。通过简化报文封装、增强安全功能和提高可扩展性，提供适用于大规模数据中心的高效、安全加密解决方案，主要包括以下要素：

1）安全域

安全域是可以进行保密和真实通信的终端集合，使用KDF公共对称密钥来保护通信。安全域内的源被分配唯一源标识符或源IP地址，每个安全域应分配一个在可达网络内唯一的安全域标识。当一个安全源加入一个安全域时，安全域管理会向安全源分配一个安全标识和提供安全参数。

2）安全域密钥库

安全域由数据包中携带的标识信息进行安全标识，而安全域标识可用于从安全域密钥库查找安全域的上下文。安全域密钥库基于报文输入参数，进行密钥解析处理，获取密钥信息，生成加密密钥。安全域密钥库安全参数管理包括：KDF模式，安全加密算法，域相关安全密钥等信息。

3）密钥分发

安全域管理实体负责将安全参数(如安全域标识，安全域密钥等)分发给安全域中的FEP，实现用户通信的安全传输。

4）加密和认证

使用带有身份认证的加密算法(AEAD)，默认采用AES-GCM，以实现对用户身份的对称加密。在完成数据的正确加密及认证后，数据包正常转发，未经加密的明文不可以在网络上发送。对接收到的数据包进行解密，并进行身份认证，认证通过后正常转发报文，认证不通过作丢弃处理。

随着AI和HPC在大规模数据中心中的应用日益增加，动态数据存储和维护的需求也在不断上升。与此同时，网络威胁的演变使得现有安全协议面临新的挑战，尤其是量子计算技术的崛起。为此，未来的安全协议需要与量子技术相结合，以确保数据传输的长期保密性。除

量子技术，未来的安全协议也可能结合如区块链、机器学习和人工智能等新技术，在增强数据保护和威胁检测能力同时，提高在不同应用场景应用的适用性和灵活性。

6 高性能广域网关键技术

6.1 主动拥塞避免

HP-WAN场景中，并发流传输将导致网络拥塞，而由于网络不感知业务，对流量传输的带宽资源无法规划和预留，可能导致突发大面积拥塞，且传统端侧调速的拥塞控制方法导致带宽利用率下降，导致现有的拥塞机制无法满足高通量传输要求，需要依赖网侧提供主动拥塞避免机制，主动降低丢包和时延。主动拥塞避免机制包括基于配额的流控，基于时隙化的队列管理等，可提供长距无损或者一定的确定性能力。

6.2 近源端反馈机制

HP-WAN 的长距离传输带来极长的链路传输延迟和较大的 RTT，将导致网络状态反馈延迟，端侧无法及时调整传输速率。例如将 ECN 技术应用于 HP-WAN 时，长距下端侧收到服务端 ECN 反馈的时延变大，慢反馈可能导致导致端侧调速不及时，缓存溢出。网络需要引入基于 offered load 的快速反馈机制，在近源端及时反馈。

6.3 端网协同速率协商

在HP-WAN场景中，突发的流量数据传输可能会导致网络内的瞬时拥塞、丢包和排队延迟，由于在拥塞控制机制中，端侧对网络的带宽资源无量化感知，导致调速不平滑，吞吐量下降。因此，初始速率协商是网络的重要组成部分，决定数据的起始速率，如果初始速率设置得太低，可能会导致带宽利用率不足，无法完全利用网络的潜力，如果设置得太高，可能会导致网络拥塞，导致数据包丢失和传输增加延迟。为了确保高效的高通量，需要提供端网速率协商机制，端网协同实现快速拥塞控制。

6.4 任务式传输及配额调度

在HP-WAN场景中，数据传输有任务式传输的需求，且任务有预期性，需要提供任务感知及资源调度，保障所有任务的传输需求及资源保障。基于配额（quota）的调度是一种资源管理策略，配额可定义为一定时间内的可用资源（带宽，队列，buffer等），包括动态

和静态资源，静态就是申请资源预留的，比如确定性链路，没有申请资源预留的，可使用共性资源。网络入口可以进行基于配额的速率协商和流量准入，控制器需要进行基于配额的资源规划、分配等。

6.5 基于流的网络监控

当HP-WAN数据传输中发生故障时，需要找出故障原因和识别丢包或拥塞节点。Telemetry可以用于HP-WAN提供基于流的网络监控，对网络带宽、流量和性能等进行自动化管理和监控。例如，带宽监控对于HP-WAN网络规划和配额保障非常重要，网络运营商可以预测带宽可用性并保证高通量传输。性能监控是管理和优化网络的关键，如端到端或逐跳节点丢包、延迟、抖动、跳数、队列和buffer信息等，可以用于优化HP-WAN网络拥塞，提升吞吐量及带宽利用率等，保障HP-WAN高通量传输。

7 展望

2024年，OpenAI发布的O1模型预示着AI进入了新一轮技术高速发展周期，而刚刚发布的Grok-3训练集群已达到20万卡级别，Scaling Law不仅没有停滞，对算力规模的需求还在持续，只是重心从预训练向后训练和推理转移。AI集群规模从万卡向十万卡甚至百万卡演进已经成为业界关注的焦点，作为AI基础设施重要组成部分的高性能网络也将迎来新一轮技术革新，如下几个技术趋势已经相当明朗：

趋势一：以太网成为Scale Out和Scale Up网络创新的基础

Scale Out网络面向超大规模设计，而Scale Up网络面向极致性能设计，这必然导致两者在带宽、时延以及传输效率方面呈现出显著差异。但抛开上层协议设计细节不谈，将以太网作为两类网络共同的物理层基础已经成为行业共识，传统的PCIe和Infiniband都不再是未来大规模智算集群网络的主流选择。

趋势二：光电融合是下一代智算集群网络发展的必然方向

受摩尔定律和香农定律的双重约束，当前电互联技术已经进入发展的瓶颈期，下一代智算集群对性能和能耗的极致追求必然导致光互联技术得以广泛应用。具体而言，在Scale Up网络域，以OIO/CPO为代表的光互联技术将为芯片间、板内和板间互联带来带宽和功耗的巨大提升空间；在Scale Out网络域，CPO以及OCS光交换技术的引入，在确保智算集群规模的前提下，为网络提供更高的能效比和更灵活的组网能力，光互联和电互联技术的深度融合将成为下一代智算集群网络发展的必然方向。

趋势三：可重构能力会成为下一代智算集群网络的基本要求

除了规模本身，AI大模型自身的发展也存在一定的不可预测性，模型的高速迭代和AI基础设施漫长的建设周期之间不匹配的问题将更为突出，下一代智算集群网络可能需要支撑几代需求迥异的大模型训练或者推理，“One Network for All”的前提可能不再成立，可重构能力会成为下一代智算集群网络的基本要求。具体而言，未来网络需要通过软硬件的协同创新，在流量工程、物理拓扑、多租户隔离等方面具备一定的可按需重配甚至拓扑重构的能力，从而实现“Demand Oblivious（需求无感）”向“Demand Aware（需求感知）”网络的转变。

趋势四：AI负载对广域网的长期影响被低估，广域网将迎来一次架构升级的周期

目前业界对智算网络的研究更多聚集在数据中心内，但AI业务的蓬勃发展对运营商广域网络的中长期影响被严重低估，超大型智算集群不可避免的跨DC协同训练、AI训练对数据量和流转效率的迫切需求、个性化AI内容让传统CDN机制失效、AI推理资源的分布式下沉等诸多因素讲是未来广域网络需要面对的变量，广域网将不仅仅只是带宽的升级，而是数据中心高性能网络的广域延伸，不久的将来或将迎来一下架构升级的周期。事实上，根据Meta最新公开的数据，在AI业务的驱动下，2023-2024年间Meta骨干网流量呈现出30%以上的高增长，且AI流量绝对值已经超过了传统流量。

趋势五：高性能专线与广域公网并存，高通量广域公网受业务和成本双重驱动

受在线和线下计算两种不同场景的差异化需求驱动，高性能广域网将在可预见的未来并存专线和增强性公网两种模式，前者面向时延极度敏感的在线分布式训练等场景，后者面向传输通量敏感的线下业务。受可扩展性及成本等多重因素驱动，增强性广域公网长期来看，将成为市场的自然选择。在传统端侧为主的技术机制基础上，广域网侧将以更加主动和精细化的模式介入端到端高性能传输流程中，从而根本性解决高性能广域网的瓶颈问题。

8 参考文献

[1]. Hu Y, Eran H, Firestone D, et al. Congestion control for large-scale RDMA deployments[J]. ACM SIGCOMM Computer Communication Review, 2015, 45(4): 523-536.

[2]. Hoefler T, Roweth D, Underwood K, et al. Datacenter ethernet and RDMA: Issues at hyperscale[J]. IEEE Computer, 3, 56(7): 67-77.

[3]. Chaturvedi S K. Network reliability: Measures and evaluation[M]. Hoboken: John Wiley & Sons, 2016.

[4]. Zhu J, Huang N, Wang J, et al. Availability model for data center networks with dynamic migration and multiple traffic flows[J]. IEEE Transactions on Network and Service Management, 2023, 20(3): 2975-2989.

[5]. Brown M. MAC link latency considerations[EB/OL]. 2023-04-13. https://www.ieee802.org/3/dj/public/adhoc/optics/0423_OPTX/brown_3dj_optx_01b_230413.pdf.

[6]. IEEE. IEEE Standard for Ethernet: IEEE 802.3-2022[S]. New York: IEEE, 2022-07-29.

[7]. IEEE P802.3bs 200 Gb/s and 400 Gb/s Ethernet Task Force. Objectives[EB/OL]. 2017-03-16. https://www.ieee802.org/3/bs/Objectives_16_0317.pdf.

[8]. Kang Y, Wang X, McGlohon N, et al. Modeling and analysis of application interference on dragonfly+[C]//ACM SIGSIM Conference on Principles of Advanced Discrete Simulation. New York: ACM, 2019: 161-172.

[9]. RFC8279. Multicast Using Bit Index Explicit Replication (BIER)[S]

[10]. RFC9197. Data Fields for In Situ Operations, Administration, and Maintenance (IOAM)[S]

[11]. Chen, X., et al. A Framework and Definition for Collective Communication Offloading[EB/OL]. 2024-07-08. https://datatracker.ietf.org/doc/draft-chen-rtgwg-cco-framework-and-definition/.

[12]. Ramakrishnan, K., et al. The Addition of Explicit Congestion Notification (ECN) to IP[EB/OL]. 2001-09-01. https://datatracker.ietf.org/doc/html/rfc3168.

[13]. IEEE. IEEE 802.1Qbb. Priority-based Flow Control[EB/OL]. 2020-06-07. https://1.ieee802.org/dcb/802-1qbb/.

[14]. P4. INT v2.1[EB/OL]. 2020-11-11. https://p4.org/p4-spec/docs/INT_v2_1.pdf.

[15]. Kumar, A., et al. Inband Flow Analyzer[EB/OL]. 2024-04-26. https://datatracker.ietf.org/doc/draft-kumar-ippm-ifa/.

[16]. Ravi, S., et al. Congestion Signaling (CSIG)[EB/OL]. 2024-02-02. https://datatracker.ietf.org/doc/draft-ravi-ippm-csig/.

[17]. Nguyen, E., et al. Poseidon: Efficient, Robust, and Practical Datacenter CC via Deployable INT[EB/OL]. 2023-12-15. https://www.cs.rice.edu/~eugeneng/papers/NSDI23.pdf.

[18]. Miao, R., et al. HPCC++: Enhanced High Precision Congestion Control[EB/OL]. 2024-02-29. https://datatracker.ietf.org/doc/draft-miao-ccwg-hpcc/.

[19]. Xiong, Q., et al. Problem Statement for High Performance Wide Area Networks[EB/OL]. 2024-12-05. https://datatracker.ietf.org/doc/draft-xiong-hpwan-problem-statement/.

[20]. Cardwell, N., et al. BBR Congestion Control[EB/OL]. 2024-10-21. https://datatracker.ietf.org/doc/draft-ietf-ccwg-bbr/.