中国电信
CHINA TELECOM

智算网络系列技术白皮书

分布式智算中心无损网络技术白皮书

中国电信股份有限公司研究院 I 2024.8

版权声明

编写说明

主要编写单位：

中国电信股份有限公司研究院、中国电信股份有限公司北京分公司

主要编写人员（排序不分先后）：

傅志仁、雷波、顾鹏、叶平、王江龙、李聪、解云鹏、王学聪、李云鹤、冀思伟、刘宇旸、吴楠、张越、马小婷、周舸帆、唐静、王轶、张勇

高级顾问（排序不分先后）：

张文强（中国电信集团公司）

罗锐（中国电信北京分公司）

史凡（中国电信集团公司）

胡芳龙（中国电信集团公司）

撰写团队联系方式：

中国电信股份有限公司研究院

解云鹏

010-50902166

xieyp6@chinatelecom.cn

前　言

2024 年 3 月，政府工作报告中首次提出开展“人工智能+”行动，打造具有国际竞争力的数字产业集群。这意味国家将加强顶层设计，加快形成以人工智能为引擎的新质生产力。随着这一行动的深入推进，人工智能将在推动产业升级、促进新质生产力加快形成等方面发挥重要作用。

随着人工智能的浪潮来袭，以大模型为代表的 AI 方案逐步深入千行百业，算力需求日益攀升，智算基础设施的重要性进一步凸显。然而，在智算基础设施建设过程中尚面临组网、通信、能耗、成本等多重挑战，行业要“以网强算”，通过无处不在的网络资源，补齐单点算力规模不足的差距，夯实智算业务发展基础。

本白皮书聚焦 AI 大模型下智算业务的典型需求和特征，对分布式智算中心无损网络方案、核心技术展开深入研究，并积极推动分布式智算中心互联现网验证。我们希望通过白皮书的研究与分析，得到更多同行的参与和讨论，同时也期盼与众多合作伙伴一起携手并进，汇聚行业力量，共同打造大规模、高带宽、高性能以及智能化的 AI 大模型分布式智算中心网络。

目　　录

1 分布式智算中心无损网络场景及需求

1.1 智算业务的典型需求和特征

从 Transformer 问世至 2023 年 ChatGPT 爆火，人们逐渐意识到随着模型参数规模增加，模型训练的效果越来越好，且两者之间符合 Scaling law 规律。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能力迅速提升。例如，拥有 1.8 万亿参数的 GPT-4 在复杂问题的处理能力方面远超 GPT-3，谷歌的 Gemini 大模型性能也超越其早期版本。但提升模型参数的规模和性能后，AI 大模型训练对于网络的需求也会发生巨大变化。

在大模型训练场景下，随着参数规模从亿级提升到万亿级别，算力需求呈现“爆发式”增长。据统计，2012~2022 年模型算力需求每年增长 4 倍，而 2023 年后模型的算力需求以每年 10 倍的速度增长。这意味着训练超大 AI 模型需要数千/万卡 GPU 组成的集群高速互联。此外，机内 GPU 通信和机外集合通信将产生大量通信需求。例如，千亿级参数的大模型并行训练所产生的集合通信数据将达到数百 GB 量级。若要在极短时间内完成参数交换，将对 GPU 与 GPU 间、GPU 与网卡间、网卡与网卡间的超高带宽互联提出较高要求。网络拥塞和丢包也会严重影响 GPU 计算效率，据实验统计，0.1%的网络丢包率就会带来 50%的算力损失，因此提升通信性能可有效释放智能算力。

AI 大模型训练/推理需要智算网络具备超大规模、超高带宽、超低时延、超高可靠等关键特征。如何设计高效的集群组网方案，提升 GPU 有效计算时间占比（GPU 计算时间/整体训练时间），对于 AI 集群训练效率的提升至关重要。

1.2 分布式智算中心无损网络场景

超大规模 GPU 集群成为大模型训练的必要条件，而算力需求的指数级增长对 AI 基础设施带来极大挑战。在构建万卡甚至十万卡集群时，由于机房空间/电力不足、机房散热等问题，智算中心单点算力规模建设受限。

为破解智算基础设施供给难题，中国电信践行“以网强算”的技术路线，即利用无处不在的网络资源弥补小规模智能计算的差距，再结合集中式的算力调度

策略，提升整网智算利用率。目前，“以网强算”已成为国际格局和产业环境下中国最具优势的发力点。

“以网强算”将多个智算中心互联成一个大型虚拟智算集群，通过分布式智算中心无损网络（也称 RDMA 拉远），实现区域内多智算中心协同计算，满足更大规模的算力需求。目前，分布式智算中心无损网络主要适用于两类场景：算-算拉远和存-算拉远。

（1）算-算拉远场景

我国单点智算中心规模普遍偏小，规模为 100-300PFLOPS 的小型智算中心占比超 70%，而规模超过 1EFLOPS 的大型智算中心仅占 25%，且多由云提供商及大型企业自建，集中在京津冀、长三角和粤港澳。算-算拉远可以将区域内多个已经建成的智算中心的算力进行整合，从而无需建设超大规模集约型智算中心就能够训练更大的模型。

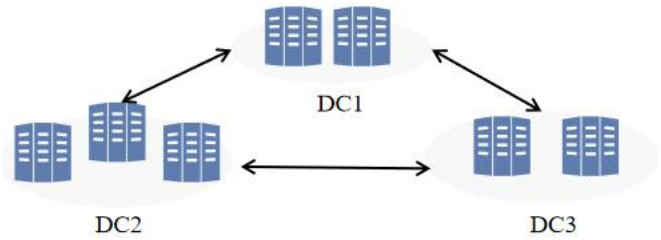


图 1-1 多智算中心合一场景

此外，单个智算节点往往会存在资源利用率不足、闲散算力资源浪费的问题。在算力使用过程中，租户算力诉求与实际部署算力往往不一致，导致算力零散在本地，智算中心算力资源碎片化。如何把零散的资源整合起来，系统优化算力基础设施，布局盘活机房，促进跨集群算力高效互补和协同联动成为充分发挥算力的关键能力。算-算拉远能够充分利用碎片资源来执行合适的任务，提升系统利用率。

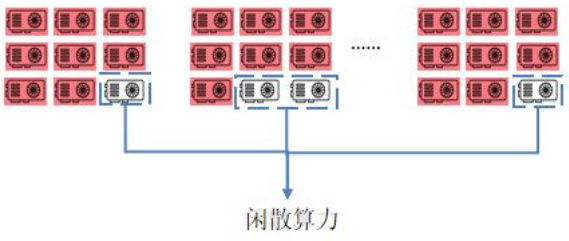


图 1-2 碎片资源整合场景

（2）存-算拉远场景

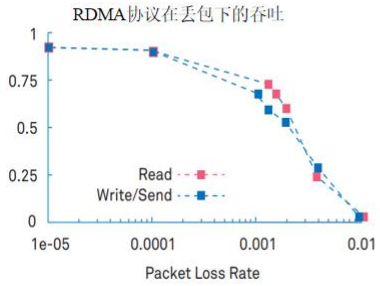
高性能、高可靠存储是公有云最基础的服务之一。当前公有云中广泛采用存算分离架构，即计算集群和存储集群可能位于 Region 内的不同 DC 中，而互连计算集群和存储集群的网络成为实现云存储服务高性能和高可靠性的关键。存-算拉远可以将 Region 内的计算集群和存储集群无损互联，满足数据本地化需求，保障数据安全。

1.3 分布式智算中心无损网络挑战

在探索跨智算中心构建超大规模智算集群过程中，算力和网络均遇到了诸多问题和挑战。首先，集群拉远部署相比于本地集群部署在 DCN 协议面需要解决时延和丢包两个难题。

（1）拉远增加网络传输时延：AI 训练每轮迭代会通过集合通信进行参数同步，而集合通信内部存在多轮数据交互，以及多次跨长距通信。长距拉远后，传输距离每增加 10km，通信时延增加 10ms 左右，对 AI 大模型的训练效率产生极大影响。

（2）网络拥塞丢包，使性能急剧下降：当前 AI 训练采用 RDMA 协议，而 RDMA 的高效率依赖于极低的丢包率。数据显示，当网络的丢包率大于 10^{-3} 时，RDMA 有效吞吐将急剧下降；2%的丢包率会使 RDMA 吞吐率下降为 0。因此，要使得 RDMA 吞吐不受影响，丢包率必须保证在十万分之一以下，最好为零丢包。在长距拉远场景下，当网络出现拥塞时，若没有在 RTT（往返时间）内及时缓解拥塞，就会发生丢包，导致一轮迭代训练时间增加，大模型的训练效率下降。



拉远距离	120km	
是否丢包	否	是
一轮迭代训练时间（ms）	7423.9	13994.6

图 1-3 丢包影响 RDMA 吞吐

其次，集群拉远部署和本地集群部署相比在传输网也需要解决高带宽和稳定性难题。

（1）**超大带宽、灵活组网保证长距拉远算效：**在跨 DC 分布式训练场景中，需要提供充足的互联带宽，并根据智算中心空闲服务器数量灵活组网，避免网络拥塞，实现高效传输。

（2）**高可靠机制保证 AI 训练的稳定：**检查点（checkpoint）机制是 AI 训练的必要需求，主要用于在训练过程中保存模型的权重，以便在训练中断或模型更新时恢复训练，从而提高训练的效率和稳定性。网络还需要具备抗多次断纤能力，防止网络故障引起 AI 训练中断。

（3）**故障分钟级检测及定位：**模型训练期间可能受施工震动、挤压弯折、意外挖断、接头松动、老化等影响，从而导致光缆故障，训练也会随之中断。为保证训练的稳定，要求网络具备故障时分钟级自动检测和定位、分钟级提前预警的能力，以保证智算拉远训练时的高可用。

针对以上难题，若要实现长距无损传输，需要协同优化 IP 层和光传输层技术，构建分布式智算中心无损网络，实现多数据中心协同提供服务。在 IP 层，一方面可以优化集合通信算法，减少长距链路的流量传输，从而消除流量交叠现象；另一方面可以引入全局负载均衡和精准流控技术，实现多节点互联网络的无拥塞、高吞吐。在光传输层，一方面可以依托城域网或区域网延伸覆盖智算节点，并在资源不足区域新建 800G/1.2T 超大带宽的互联网络，构建高品质光互联；另一方面，可以提高网络故障处理能力，实现高可靠、智慧化运维。

1.4 业界研究概况

大模型推动智算基础设施建设快速发展，但电力供应、机房空间成为大规模智算建设的瓶颈。业界正在积极探索将分布在多个智算中心的算力协同起来，进行跨 DC 的大模型分布式训练。

谷歌利用自研低成本、高性能 TPUv4 超级计算机（SuperPod）满足大模型训练/推理算力需求，其中每一个 SuperPod 可以提供 1 Exaflop 级（每秒百亿亿次浮点运算）的运算能力。目前，谷歌已经部署了数十台 TPUv4 SuperPod，并完成跨多个数据中心的 Gemini Ultra 大模型训练，此前 5400 亿参数语言模型 PaLM

也是用 2 个 TPUv4 SuperPod 训练的。OpenAI 与微软也在计划建设十万甚至百万级 GPU 卡的算力集群，以满足 GPT-6 模型训练需求。但由于电力受限，预计将 GPU 卡分布在几个或几十个地区，并利用开放 Ethernet 协议替换 IB 协议来实现跨区域 GPU 之间的互联。Meta 宣布推出两个具备 2.4 万个 GPU 卡的 AI 集群，分别采用 RoCE 和 IB 协议，并在硬件、网络、存储、性能等方面进行深度优化，以支持大语言模型如 Llama 3 的训练。为了解决 AI 训练集群造价昂贵问题，Meta 又提出去中心化异构训练，利用分布式、异构和低带宽互联的 AI 训练资源来训练基础大模型，降低训练成本。

阿里提出“双上联+双平面+多轨”的 HPN7.0 网络架构，该网络架构中单个 Pod 规模已经达到 15K GPU，可满足绝大多数 LLM 的训练需求。为建设更大规模智算集群，设计了不同 Pod 之间通过核心层互连，从而在单个集群中支持超过十万个 GPU 节点。目前，HPN7.0 网络架构已经在阿里云上线运行 9 个多月，实践表明 LLM 训练的吞吐性能相比传统数据中心网络而言提升了 14.9%。百度智能云基于 CENI 打造了跨广域工业视觉大模型算网融合技术，依托百度自研昆仑芯以及百度文心视觉大模型，将算力、网络、大模型和应用场景融合，实现行业大模型跨广域精调、推理服务。测试结果显示，在相距超 300 公里的两地之间，该技术使行业大模型跨广域推理效率提升 42%。此外，NTT 在 Mitaka 和 Yokosuka 之间通过全光子网络（APN）搭建 LLM 远程训练测试环境，将训练数据保存在企业本地，而使用数百公里外数据中心的 GPU 进行训练，训练效果与本地的训练效果相近。

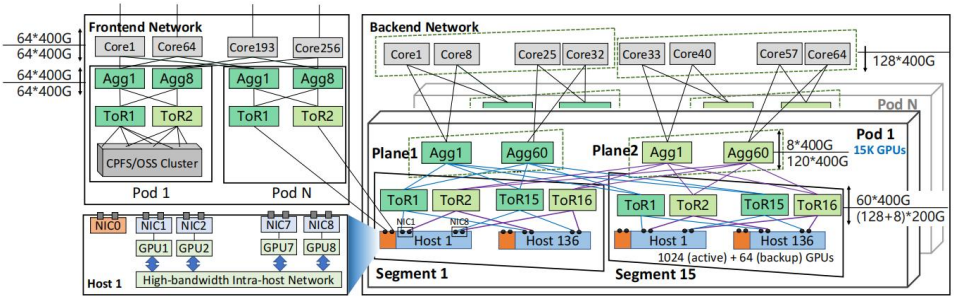


图 1-4 阿里 HPN7.0 架构

当前单点智算中心算力规模受限、算力资源碎片化严重，难以承载大规模 AI 训练业务。采用跨 AZ、跨 Region 的多个数据中心组成的 AI 训练集群可有效

支撑十万卡甚至百万卡级别的 AI 训练任务，同时提高资源利用率，是未来智算产业发展和探索的重要方向。

2 分布式智算中心无损网络解决方案设计

2.1 方案设计原则

分布式智算中心无损网络是一种特别设计的网络架构，通过全栈创新，旨在整合盘活闲散算力资源，实现算力高效互补和联动，进而构建极致可靠的算力集群，为大规模分布式智能计算提供高性能、低延迟且无丢包的数据传输能力。这种网络架构可以提供接近于本地智算中心网络性能的计算效率和数据处理速度，对于支持大规模机器学习模型训练和高性能计算至关重要。

分布式智算中心无损网络在方案设计时，应遵循打造超大规模算力集群、提供高效稳定训练能力、实现算网灵活调度供给以及坚持绿色低碳节能减排四大设计原则：

（1）打造超大规模算力集群

当前智算集群主要规模为单数据中心内的数千张计算卡，更大规模的万卡乃至超万卡集群建设尚处于初期阶段。构建超大规模算力集群将进一步缩短大模型训练时间，加速模型能力迭代。通过分布式智算中心无损网络可以实现多节点算力协同，构筑超大规模的极致算力集群。

（2）提供高效稳定训练能力

大模型的计算量大、训练时间长，训练期间涉及节点间的频繁交互，对网络稳定性要求高。如果训练期间网络出现不稳定，轻则将回退至上一个分布式训练的断点，重则可能要从 0 开始，会影响整个训练任务进度，给客户带来重大损失。分布式智算中心无损网络需要在支持大模型高效训练的同时，保持长期训练的稳定性。

（3）实现算网灵活调度供给

构建多 DC 算力集群灵活调度，实现算力高效互补和联动。同时，通过应用服务、算力使能平台和算力底座的深度适配，高效的算网调度及协同，实现训练资源的按需分配，为用户提供接近本地训练的算力效率和灵活的算力供给能力。

（4）坚持绿色低碳节能减排

通过分布式部署的算力集群分担电力，实现电力与算力的最优配置，并通过800G C+L 构建低时延、高带宽的全光网络，为智算集群提供超大带宽的主干道，实现最优成本的 bit 传输和算力的绿色供给。

2.2 分布式智算中心无损网络总体架构

分布式智算中心无损网络总体架构由多个单节点智算中心网络组成，其中每个单节点智算中心网络均包括多个业务区块：AI 集群区、通用计算区、存储区、带外管理区、管理区、网络服务区、接入区。每个区域负责特定的功能，区块间通过核心交换区的核心交换机连接在一起构成单节点智算中心网络，多个单节点智算中心网络之间通过广域互联区互联，构成分布式智算中心无损网络，共同支撑起整个分布式智算中心的运行。

分布式智算中心无损网络总体架构如图 2-1 所示：

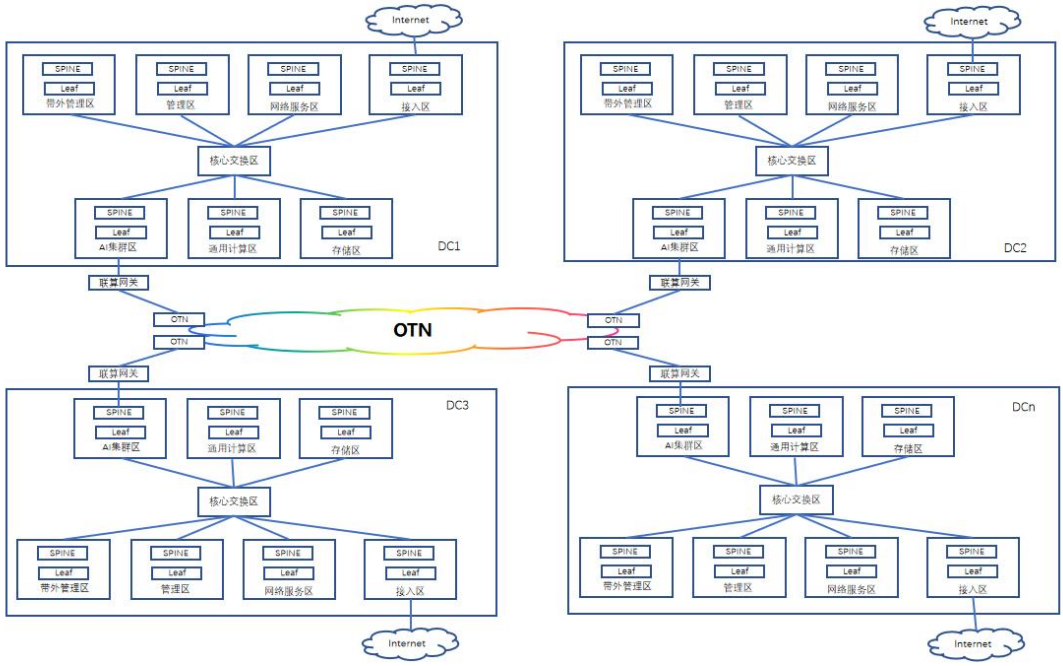


图 2-1　分布式智算中心无损网络总体架构

AI 集群区：包括 GPU、TPU 或其他加速器等高性能计算节点，用于智算集群分布式训练时的参数交换。要求网络具备大带宽、高吞吐、无丢包能力，需要部署无损网络。

通用计算区：包括 CPU 等通用服务器，支持各种类型的应用程序和服务。提供标准的计算资源，用于运行非 AI 相关的计算任务，通常部署为 TCP/IP 有损网络。

存储区：包括高速缓存存储、块存储、对象存储等多种存储类型，用于存储大量数据和模型文件。要求网络具备高速大带宽互联能力，可按需部署无损网络。

管理区：包括监控系统、配置管理系统和安全控制系统。负责整体网络的监控、配置和安全管理，通常部署为 TCP/IP 有损网络。

带外管理区：用于管理计算节点和其他网络设备的带外接口。提供独立于主网络之外的管理通道，确保即使在主网络出现问题时也能进行设备管理，通常部署为 TCP/IP 有损网络。

网络服务区：提供防火墙、负载均衡、DNS、NTP 等网络服务，保障网络设备和服务的正常运行，通常部署为 TCP/IP 有损网络。

接入区：是智算中心对外连接的主要入口。包括防火墙、负载均衡器等设备，用于连接外部网络和提供安全防护，通常部署为 TCP/IP 有损网络。

广域互联区：包括路由器、OTN 等设备。多节点智算中心通过具备高通量的联算网关互联，中间通过 OTN 全光网络提供高品质的大带宽连接，实现 AI 集群训练网络的跨 DC 互联互通，需要部署无损网络。

这些区域共同构成了分布式智算中心网络架构，每个区域都承担着特定的角色，通过相互协作确保整个分布式智算中心的高效运作。其中，构建 AI 集群之间的无损广域互联网络是方案中的设计重点。通过提供物理隔离、全程资源独享的高质量、低时延的波长级大带宽管道，实现 DC 间的多方向任意互联，并提供抗多次断纤的能力，保证互联的可靠性。在大带宽的传输资源基础上，智算中心出口通过联算网关提供灵活的、易扩展的跨智算中心组网和长距无损、高吞吐、高可靠的数据承载。AI 集群区网络互联架构如图 2-2 所示：

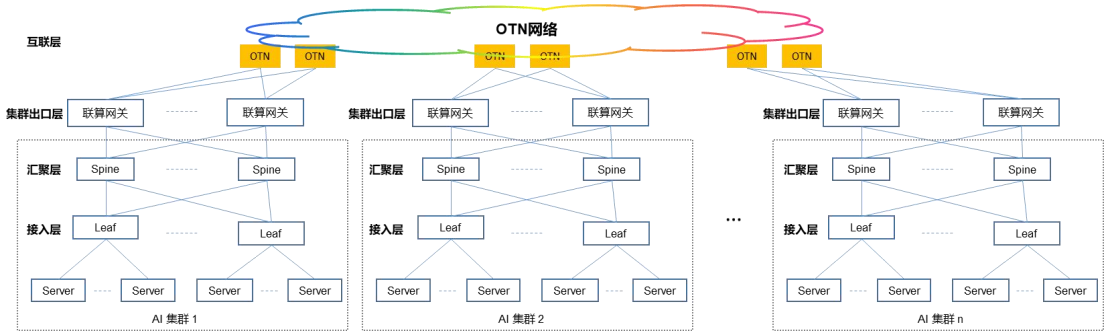


图 2-2　AI 集群区网络互联架构

AI 集群网络架构从下到上分成四个层次：

接入层：由 Server Leaf 交换机组成，支持 AI 算力服务器的高密规模接入，上下行带宽收敛比推荐 1:1。AI 训练服务器每个接口采用独立 IP，以独立链路方式接入到 Server Leaf 交换机，不做链路捆绑。接入侧支持光模块故障保护机制，避免接入侧链路故障导致训练中断。

汇聚层：由 Spine 交换机组成，下行接 Server Leaf 交换机，上行接 DCI Leaf 交换机。Spine 交换机的数量决定了本节点 AI 集群的总规模，根据训练业务模型的选择，汇聚层可以有一定的收敛比。

集群出口层：由联算网关组成，作为 AI 集群的出口，联算网关下行与多 Spine 交换机进行全互联，上行通过 OTN 和其他节点互联。集群出口层也可根据业务模型的选择进行收敛。此外，集群出口层采用算网协同、DC 间与 DC 内级联精准流控等技术，实现网络负载均衡和长距无损，为 AI 集群的高效训练提供基础网络保障。

广域互联层：不同智算中心节点之间采用 OTN 全光网一跳直达，全程无拥塞，无丢包。广域互联层提供单纤 96Tbps 的超大带宽能力，利用高性能的 WSON 技术和智能运维技术，保障智算高可靠互联，同时具备与业务联动的波长级拆建能力，实现算网协同。

通过这些设计，AI 集群网络架构能够在长距离、大规模的分布式计算环境中提供稳定、高效的数据传输能力，为大规模智算中心的高效运行提供坚实的基础。

2.3　方案技术特征

分布式智算中心无损网络将智算中心无损网络从数据中心网络向广域网延伸，方案具备长距无损、超大带宽、超高可靠、弹性敏捷和智慧运维的特征。

（1）长距无损：在大模型训练过程中，采用 RDMA（远程直接内存访问）作为输入输出协议。由于 RDMA 对网络拥塞和丢包非常敏感，即便是少量的丢包也会导致性能急剧下降。因此，底层网络必须具备无损传输能力，确保数据传输过程中不会出现拥塞或丢包现象，从而避免上层协议性能受损。

（2）超大带宽：超大带宽能够确保大量数据在分布式智算中心之间快速传输，加速 AI 模型的训练和推理过程。随着数据量的增加，分布式智算中心之间需要高效同步数据和模型参数，这就要求网络提供足够的吞吐量，以避免网络拥塞和性能下降。

（3）超高可靠：为了保证分布式智算中心之间的长期稳定训练，防止网络施工等外来因素导致的训练中断，传输网络需要具备高可靠性。例如在网络链路发生故障时能够快速恢复，保证智算不中断，任意二次故障带宽不下降，以避免因链路中断而导致的智算训练回退和算力效率下降。

（4）弹性敏捷：分布式智算中心无损网络需要根据多租户的不同需求，能够灵活地组建不同规模和类型的集群组网。这意味着网络需要具备弹性敏捷的按需拆建能力，能够根据计算需求的变化快速调整，动态分配大带宽资源。

（5）智慧运维：传统网络运维面临同缆&同沟、误码闪断等难题，导致保护机制失效和业务异常。分布式智算中心无损网络需要具备智慧运维能力，能够快速准确地定位和解决问题，提高故障定位的准确率，确保网络的稳定运行。

3 分布式智算中心无损网络核心技术

分布式智算中心无损网络在 IP 网络层和光传输层都需要引入新的技术点，以实现长距无损、超大带宽、超高可靠、弹性敏捷、智慧运维等需求。关键的技术点总体视图如下：

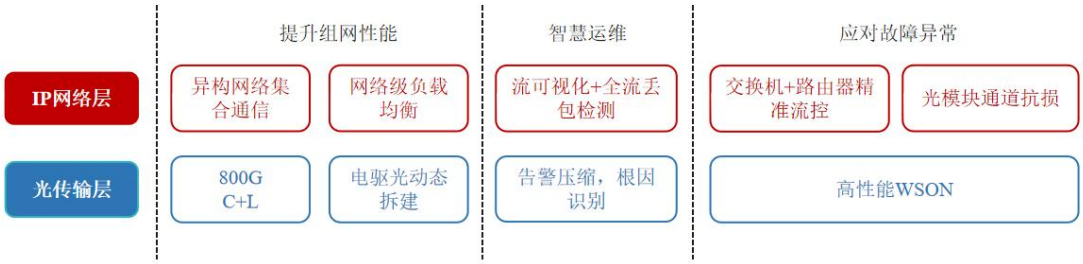


图 3-1 分布式智算中心无损网络关键技术点

3.1 异构网络集合通信优化技术

异构网络集合通信算法针对异构网络设备带宽和时延不对称（主要针对长距链路）的问题对智算业务流量进行调整，从而大幅度降低链路拥塞的可能性。在同构网络场景下，业务流量具有高度的对称性，每个节点承担的带宽业务压力是相同的。而在异构网络场景下，网络设备的处理能力不同，因此业务流量也需要调整以适应新的网络情景。例如减少长距链路上传输的数据量和传输次数，从而大幅降低长距链路拥塞的可能性。

智算业务的通信模式为集合通信，其中最主要的是 AllGather 和 AllReduce 集合通信。集合通信的特点是所有主机都会进行相同的操作，如图 3-2 所示。

AllGather：多台主机把数据的不同部分发给所有主机。

AllReduce：多台主机把数据的相同部分发给所有主机。然后所有目的主机都做一定的操作，例如求和、求最大值、求平均。

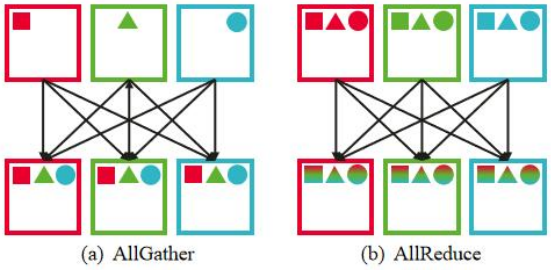


图 3-2 集合通信操作

针对这两种集合通信，业界主流的集合通信算法包括 Ring 算法和 Halving-Doubling（HD）算法。其中 Ring 算法通信模式简单，每台主机只需跟自己的邻居通信；HD 算法通信模式较复杂，但通信次数比 Ring 算法少，静态时延带来的开销小，因此对于小字节的通信效果更佳。然而，无论是 Ring 还是 HD，都是针对完全同构的系统设计的，集合通信的每个 Rank 行为一致，收发流量也一致。

在长距拉远场景下，网络不再同构，跨长距的 GPU 通信时延要显著高于 DC 内的 GPU 通信时延，因此传统算法将不再最优。下表总结了 Ring 算法和 HD 算法在拉远场景下的跨长距通信次数和通信量。其中 S 是集合通信数据量，N 是参与集合通信的 GPU 数量。

表 1 典型集合通信算法跨长距性能评估

集合通信算法	跨长距通信次数	跨长距通信数据量
Ring	~2N	~2S
Halving Doubling	$2\log_2 N$	NS

理想情况下，跨长距只需要进行一次通信，并且传输的数据量为 S 即可。基于该思路，设计出针对长距异构组网的集合通信算法框架，如图 3-3 所示。新算法具体步骤如下：

（1）将拉远 DC 当做两个独立的子系统，在每个 DC 内先进行集合通信操作，集合通信算法可选用 Ring 或者 HD。

（2）DC 内同步后，在每个 DC 中选取一个或者多个代表主机，然后对应的代表主机之间同步数据。例如选取 K 个代表主机($K < N/2$），则每个主机需传输 S/K 的数据。这一步的通信在网络上就是 K 个点对点双向通信。

（3）每个代表主机接收到对方的数据后，进行本地加和，再将加和后的结果在本 DC 内广播/All Gather 分发出去。这样就实现了两个 DC 之间的 AllReduce 操作。

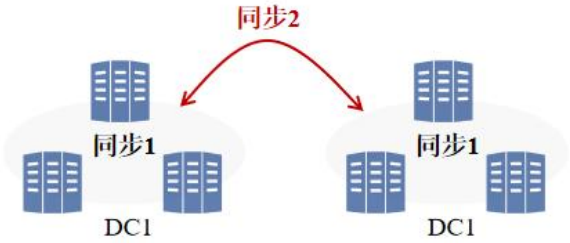


图 3-3 跨长距集合通信算法架构

图 3-4 仿真了 S=1GB 时的 AllReduce 集合通信。在拉远 100km 下，新算法相比传统 Ring 算法的性能有所提升，且随着规模增加，性能从 5%提升到 60%以上。新算法只经过一次跨长距通信，且跨长距通信数据量只有 S，均达到理论最优值。

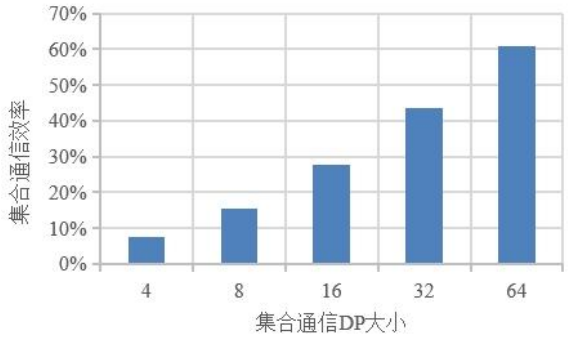


图 3-4 新算法性能仿真

在实际部署时，上述集合通信算法还需要结合网络设备来通告拓扑信息。具体来说，网络设备定期在链路层主动探测距离，构建并维护拓扑图，该拓扑图通过控制器下发到每台服务器的集合通信库。在每次执行集合通信时，根据拓扑图得到每个源端和目的端的距离，随之运行搜索算法，找到效率最高的集合通信方式。

3.2 网络级负载均衡技术

网络级负载均衡主要解决智算业务场景下非故障、同构网络的拥塞丢包问题。其中智算业务限定了网络的流量模型是集合通信。同构主要指网络设备的带宽、时延具有对称性和同步性，非故障场景指网络设备不存在光模块损坏、链路闪断、慢节点等故障问题，此时网络级负载均衡技术可以完美的将流量均衡分配到不同的网络路径，从而避免流量冲突。

智算业务流具有同步性高、流量大、周期性出现等特点。同一时刻，网络里每条等价路径上都有流经过，传统基于 ECMP 哈希的负载均衡技术无法做到所有路径的完美均衡。就像把 8 个小球随机放到 8 个盒子中，每个盒子恰好有一个小球的概率是很低的，总会有一些盒子里被放入多个小球（即链路拥塞），有些盒子没有小球（即链路闲置）。

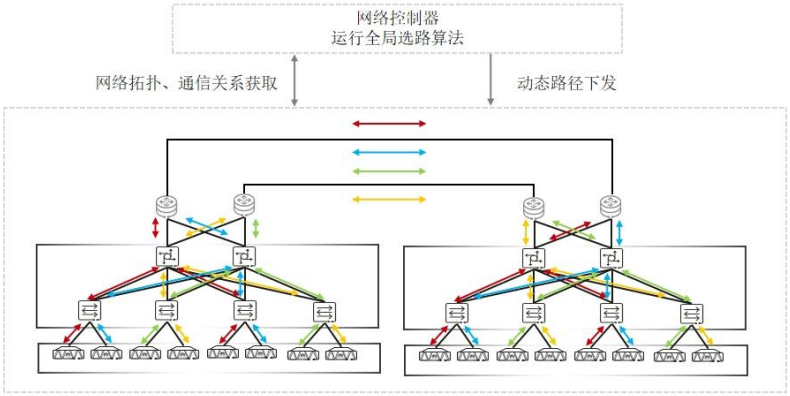


图 3-5 网络级负载均衡技术

如图 3-5 所示，网络级负载均衡技术可以通过统一规划整网流量，让所有路径之间完美均衡无冲突，避免拥塞丢包。具体来说，首先网络设备会收集业务的流量信息，并将其发给网络控制器。控制器根据拓扑、流量信息，运行全局选路算法，给每条流都选择合适的路径，做到整网完美均衡无拥塞。最后，控制器将路径信息再下发给网络设备，由网络设备作出路径调节。

3.3 精准流控技术

精准流控技术包含两种方案，一种是仅在交换机网络中使能的精准流控 1.0 方案，另一种是在跨多个智算中心时，由交换机+路由器端到端协同的精准流控 2.0 增强方案。

（1）精准流控 1.0 方案

交换机精准流控技术主要解决智算业务场景下故障丢包引起的业务性能下降问题。网络级负载均衡可以在网络正常的情况下做到整网无拥塞、无丢包。但在实际业务部署时，会出现一些异常场景，如光模块闪断、长距链路误码丢包、服务器侧拥塞导致接收数据能力下降等等。这些异常都会产生负载均衡技术难以解决的拥塞问题，进而带来异常丢包，影响训练业务性能。

在出现网络故障后，无论是链路故障还是服务器接收数据速率降低，网络有效吞吐都会下降，必然产生拥塞。但是，拥塞发生的位置不同，带来的结果也不同。如果拥塞发生在数据中心内部，因为反馈时间较短，利用流控或者拥塞控制都可以很快抑制拥塞；而如果拥塞发生在跨长距链路上，此时反馈时间变长，设备缓存不足以接纳链路在途数据包，从而发生丢包，如图 3-6 所示。

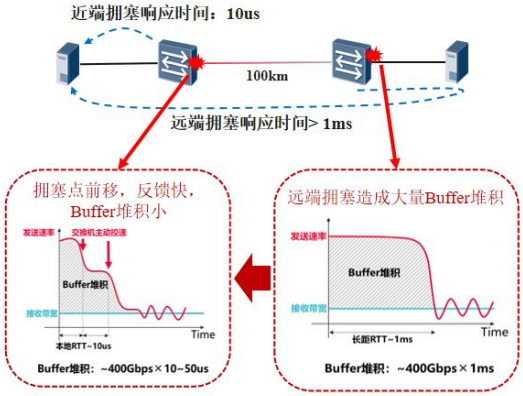


图 3-6 近端拥塞和远端拥塞带来的影响不同

交换机精准流控的思想就是当拥塞不可避免时，将原本在长距链路上的拥塞“转移”至网络第一跳设备上。具体来说，网络设备通过检查网络状态，例如端口队列堆积信息、端口反压情况，来判断是否出现拥塞。如果出现拥塞，并且该设备不是拥塞流量的第一跳设备，那么就把拥塞信息通告给拥塞流量的第一跳设备，也就是源 Leaf 交换机。随后，源 Leaf 交换机根据拥塞程度运行算法，决定以多大比例对拥塞流量进行限速。最后，源 Leaf 交换机通过发送 PFC/CNP/其它流控协议报文，实现对流量的控速，如图 3-7 所示。

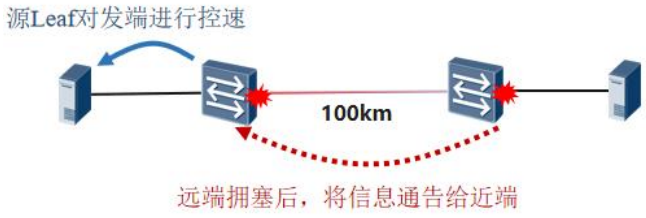


图 3-7 交换机精准流控技术

大模型训练的流量具有周期性的特点，即同一条流如果在前一个周期出现拥塞，无论是链路故障导致的流量冲突，还是目的主机接收能力下降，这个拥塞在下一个周期还会出现。基于这一特征，源 Leaf 交换机需要维护一张信息表，用于记录哪些流会发生拥塞。这样，当拥塞流后续周期性出现时，可以第一时间进

行控速，而不必再通过远端拥塞点通告后进行控速。因此，利用精准流控技术在第一周期获取到整网拥塞信息后，后面所有周期都可以做到流量无损。

（2）精准流控 2.0 增强方案

在多智算中心协同进行模型训练时，拥塞和故障可能发生在网络的任意节点或链路上，智算中心间距离的拉远会导致传输时延增加，影响网络状态反馈的及时性。路由器与交换机通过精准流控技术相互配合，不仅能够应对网络中突发的拥塞挑战，还能够在长周期故障下保障业务性能不下降。同时基于流的反压机制可以有效遏制拥塞和故障导致的反压扩散，显著提升整体网络吞吐率。

相较于传统的 PFC 机制，路由器精准流控技术解决了 PFC 的头阻、反压风暴和死锁问题，实现了从端口级流控到数据流级流控的飞跃，其基于 IP 数据报文的五元组作为流识别粒度，实现了对网络中每一条流的独立监控与动态调整，将拥塞和故障带来的影响最小化。

在跨 DC 场景下，网络环境更加复杂多变，路由器精准流控技术通过以数据流为单位的精准流量控制和精细化缓存调度，实现长距网络环境下数据的无损传输，确保数据传输的连续性和完整性。对于长周期的故障情况，路由器精准流控技术的优势更为显著。它不仅能够在故障期间通过精准流量控制，避免丢包现象，还能在限速策略上实现高度精准，包括数据流限速的开始与解除时间、限速速率的精确设定。从而确保网络吞吐能力在故障期间仍能逼近极限物理带宽，避免因限速不准导致的欠吞吐问题。

面对数据中心内高度动态的业务负载变化，路由器精准流控技术展现出极高的灵活性与智能性。其能够根据实时网络状况动态调整流控策略，实现流量峰值速率的流级别的独立控制和精准反压，有效应对网络中的突发流量，保障整体网络的平稳运行，实现故障的有效隔离不扩散。此外，路由器精准流控技术引入的弹性级联降速机制，进一步增强了网络对突发情况的适应能力，提升了网络的韧性。

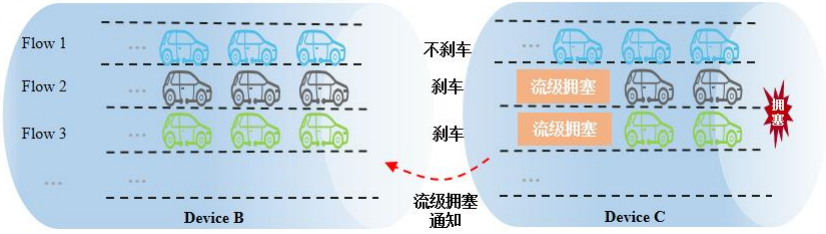


图 3-8 路由器精准流控技术

3.4 光模块通道抗损技术

网络设备间的链路故障或模块故障会导致训练中断。业界 400G/200G 光模块年失效率达 4~6‰。据统计，万卡集群平均每年发生 60 次光模块故障事件，即平均每 6 天就会发生一次。而分布式集群训练规模比单智算中心训练规模更大，面临更严峻的由光模块故障带来的训练中断问题。

大模型训练过程中，会将中间状态以 checkpoint 的形式持续记录下来，每次训练失败时不需要重头执行，而是加载最近的检查点，并继续执行。但是频繁的网络故障会使大模型训练反复回滚 checkpoint，导致整体训练效率低下。

如图 3-9 所示，激光器的失效率占比为 90%以上。200GE/400GE 短距 SR 光模块有四个通道，单激光器故障会导致整个链路故障，造成业务中断。光模块通道抗损技术可以在光模块出现单通道故障时，通过降低模块实际使用 lane 的数量，保证训练任务不中断。

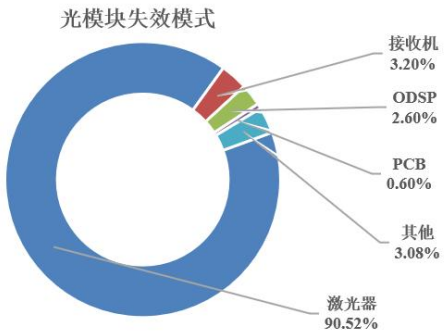


图 3-9 光模块失效模式

3.5 流可视化，全流丢包检测技术

ROCE 业务场景下，丢包会导致训练性能大幅下降。因此，智算中心内通信以及跨智算中心长距通信都对 ROCE 业务报文的传输质量提出了较高要求，希望可以做到整机全流采样、实时监控；且当链路丢包时，可以快速上传丢包发生的位置、数量和时间。管理员可以快速感知丢包，判断对网络的影响性，并及时修复故障。全流丢包监测技术支持以下能力：

（1）快速故障定位：随流检测，实时监控业务流的时延、丢包等指标；

（2）可视化：流路径可视化，网络进行集中管控。

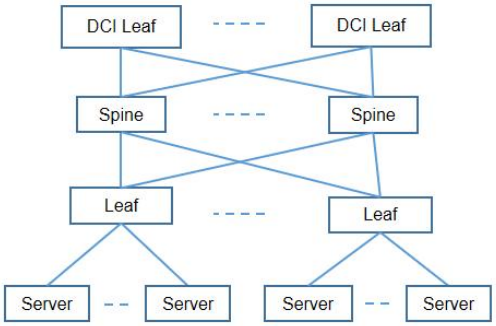


图 3-10 ROCE 业务场景

在分布式智算中心长距组网场景下，计算服务器的接入 Leaf 作为统计的 Ingress 节点和 Egress 节点，Spine 和 DCI leaf 做为 Transit 节点。

Ingress：统计流的入口测量点。Ingress 节点根据报文特征识别业务流量，对业务报文进行流标记并全流统计，统计结果发送到分析器。

Transit：统计流的中间测量点。Transit 节点识别在 Ingress 节点标记的流报文，并进行全流统计，统计结果发送到分析器。

Egress：统计流的出口测量点。Egress 节点识别 Ingress 节点标记的流报文，并进行全流统计，且在出设备时剥除流标记。统计结果发送到分析器。

全流丢包监测技术还支持丢包统计和时延统计能力：

丢包统计：在某一个统计周期内，所有进入网络的流量与离开网络的流量之差，即为承载网络在该统计周期内的丢包数。

时延统计：在某一个统计周期内，指定的两个网络节点间，同一条业务流进入网络的时间与离开网络的时间之差，即为网络在该统计周期内的时延。

3.6 大带宽传输技术

提升单端口速率可以实现超大流量的高效、低成本传输，是智算互联网络的重要发展方向之一。目前满足城域内 DC 互联的中短距 800Gbps 端口技术已经基本成熟，现已部署在智算 DCI 百公里级互联场景中，在满足智算互联百 T 级大带宽需求的同时，降低了智算互联的成本。未来需继续探索 1.2Tbps 端口速率，进一步降低单 bit 成本。

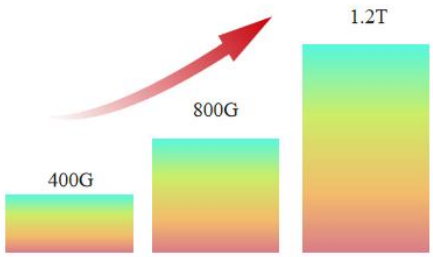


图 3-11 单播速率不断提升

随着长距传输系统由单波 400Gbps 向单波 800Gbps，甚至 1.2Tbps 演进，信号占用的谱宽不断提升。为获得更大的单纤系统容量，需要在传统 C 波段的基础上突破 L 波段相关技术，将频谱资源扩展到 C+L 波段，实现更大的单纤容量（最高可达 96Tbps 超大带宽），进一步满足智算中心之间的海量数据传输需求。

图 3-12 C+L 波段提供更大容量

3.7 波长级动态拆建技术

智算资源一般采用分时复用的方式租给不同的客户，因此，需要在任意两个算力中心之间根据空闲 GPU 数实现带宽弹性互联。网络需要匹配 GPU 数量，并根据距离、时延等不同约束，由业务侧驱动建立不同方向的波长级连接，因此 OTN 网络需要具备波长级动态快速拆建能力（简称电驱光技术）。

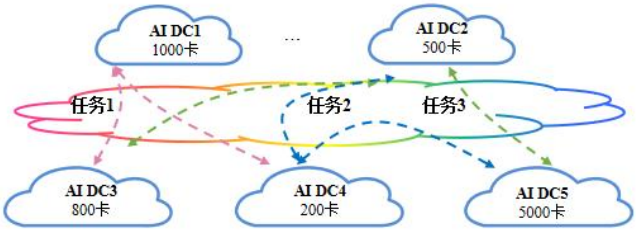


图 3-13 带宽分时复用的业务场景

电驱光技术有两种典型场景：（1）波长级的电驱光，动态拆建光层波长；（2）ODU 级的电驱光，动态拆建电层交叉+光层波长。基于录入的业务需求（例如指定源宿站点/网元、路由策略、保护等级），并结合当前网络拓扑和资源使用情况，电驱光技术可以提供以下能力：

（1）业务跨层协同算路：基于业务输入的时延和路由分离约束，自动计算满足业务需求的多条 OCH 预开通路径；

（2）光电交叉同步创建：自动生成业务配置参数，包含但不限于：Client 到 OCH、OCH 到 Fiber 的多层路由映射、波长频率与频宽配置、中继端口配置等；

（3）自动调测：基于业务跨层协同算路，自动调测开 OCH，并自动调优最佳性能状态。

3.8 高性能 WSON 技术

传统的 WSON 重路由时间为秒级到分钟级，现网测试中容易发生概率性训练中断事件，影响智算业务。因此，需要进一步提升 WSON 的重路由能力，实现确定性的光层恢复能力。当前现网重要业务采用电层 SNCP+光层重路由，通过电层 SNCP 实现 50ms 的保护能力。但在智算互联场景下，带宽为百 T 级别，电层 SNCP 要求冗余资源多，需要考虑光层的 50ms 保护能力。

针对智算百公里级互联场景，利用 WSON 50ms 技术可以在提供相同保护能力的情况下降低对资源的消耗。WSON 50ms 的关键技术包括转控分离机制、资源共享选路算法、高速报文转发技术、WSS 快速切波技术。

（1）**转控分离机制：**将路径计算、资源分配与路径建立解耦，故障时只进行路径建立所需的最少操作，避免与网络规模、业务数量的强依赖关系，提升特性应用的普适性。

（2）**资源共享选路算法：**全局统筹网络资源，并确保恢复资源可共享、零冲突且资源利用率高。

（3）**高速报文转发技术：**恢复路径建立涉及多个站点的交叉资源配置，传统方法是通过逐跳 IP 软转发实现的，但软转发实现机制与 CPU 的处理性能、重路由时的繁忙程度、协议所需传输的跳数强相关。高速报文转发技术通过使用专

有的协议报文转发芯片，可达成 ms 级的传输性能，降低了对 CPU 和业务跳数的依赖。

（4）**WSS 快速切波技术：**通过使用全新的快速液晶材料以及 LCOS 技术实现 ms 级的波长交叉切换能力。在链路故障时，WSON 可实现抗多次故障 50ms 快速恢复。

3.9 告警压缩，根因识别技术

当模型训练出现故障时，要求 10 分钟内完成恢复。因此，需要提出高效的智慧运维技术，实现分钟级的快速定位定界，防止 AI 算效长时间下降。随着 OTN 网络规模的持续增长，一个网管下面管理的网元越来越多，传统的 OTN 网络故障处理面临着更加严峻的挑战。例如：告警数量剧增会带来维护困难、根因告警识别困难、故障的定界/定位耗时费力、保护倒换等场景下因光层性能变化导致业务受影响等问题。目前，可以通过故障智能识别与余量预测来实现故障告警压缩，通过故障根因识别与 OSNR 余量精准评估来实现运维自动化。

（1）故障智能监控和识别

基于设备内生智能识别模块，将单网元内的告警标识出根因和衍生关系，上报管控系统，管控系统基于实时的告警流、现网拓扑、保护配置等，形成故障传播关系图，在线推断出故障相关的所有告警，并识别出根因告警。故障系统基于管控系统上报的智能 incident，进行跨域跨厂家的告警聚合和根因识别。

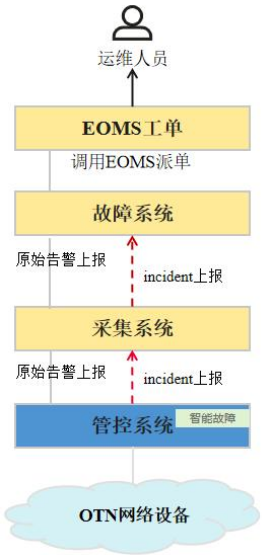


图 3-14 智能故障处理流程图

（2）性能余量智能评估

为提升光网络加掉波效率、保障业务安全，可通过对每个光波长进行数字孪生建模，提前判断系统是否可以顺利实施该运维操作且保证不会中断已有业务，从而保障加掉波、保护倒换等网络运维与优化操作的顺利完成。

通过 QoT 模型对光传输系统的物理层损伤进行精确建模，如图 3-16 所示，采用智能预测算法可以对 OCh 备路径和加掉波场景的 OSNR 余量及运维操作后的余量变化进行分析和预测，更直接准确地反映系统传输能力，并进行 OCh 劣化故障自动定界。

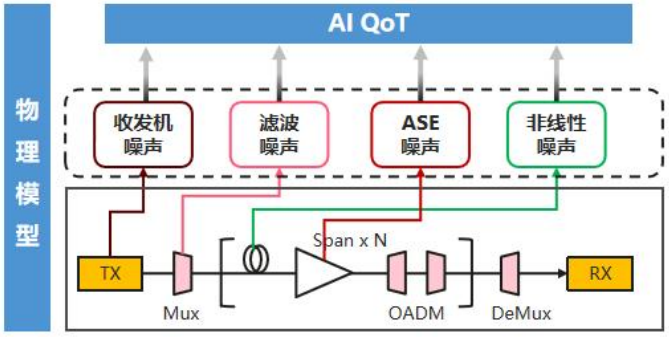


图 3-15 光网络物理层智能 QoT 模型示意图

告警压缩通过智能推理告警根因，大幅减少上报告警数量，提升现网问题定位效率；余量预测通过精准评估加掉波前的 OSNR 余量，提前预测加调波对现网波道的影响，降低了现网业务受损机率。

4 典型实践

4.1 背景与需求

2023 年以来，以大模型为代表的人工智能技术的发展已经进入了一个高速增长阶段，对经济社会发展产生了深远影响。2024 年政府工作报告明确要求开展“人工智能+”行动，打造具有国际竞争力的数字产业集群。北京数字经济发展水平位列全球第二，其中人工智能企业约 2900 家，全国占比 28%，位列第一，智算需求旺盛，是全国的智算高地。为满足未来北京市内及京津冀用算需求，以及解决单节点智算中心资源受限、不同智算中心资源使用不均衡等问题，中国电信率先在北京开展了分布式智算中心无损网络试验，验证跨数据中心合池训练的可行性，以提升区域内智算整体的供给效率。

4.2 试验概述

本试验利用 OTN 网络零丢包、低时延、大带宽的承载特点，通过全局负载均衡、长距无损流控等技术，使 RDMA 传输协议应用于广域网。目前，已在现网开展了真实场景下百公里拉远对大模型训练的影响及稳定性测试，并在全国率先完成基于高带宽、低时延的全光 800G 超高带宽传输。项目组从多拓扑、多模型、多故障等维度积极开展主流方案摸底测试，并对仿真验证结果进行分析，积极探索优化创新。

基于北京全光运力网规划，项目组先后开展了现网机房的 64 卡以及 1024 卡组网验证。一阶段在京津冀智算机房进行 80km/120km 绕行拉远验证，模拟了两个数据中心组网，组网拓扑如图 4-1 所示。二阶段在武清、瀛海、永丰三机房开展百公里分布式大模型训练，验证当前分布式智算中心无损网络解决方案在真实业务场景下的效果，并探索分布式智算集群对大模型训练性能影响的关键因素，组网拓扑如图 4-2 所示。在前期百卡、百公里拉远验证基础上，三阶段在京津冀智算机房开展了千亿参数、千卡规模 120km 两点拉远验证，组网拓扑如图 4-3 所示，本阶段探索长距链路带宽收敛情况下模型训练的性能，目标是推动无损智算互联网络的技术进一步突破。系列试验均验证了在不同拓扑

中分布式智算中心无损网络方案的有效性和稳定性。此外，模拟了多种试验中可能出现的故障情况，以验证方案在面对线路路障、服务器端口故障及其他异常情况时的韧性和恢复能力。

模型选取方面，在百卡组网规模下开展了 LLAMA2-7B、LLAMA2-13B、LLAMA2-34B、中国电信启明网络大模型-14B、Bloom-7B、Baichuan2-13B 四类百亿参数模型的分布训练验证；在千卡组网规模下进行了 Qwen-70B、GPT-175B 等模型的验证测试。通过多模型验证可以确保智算拉远方案能适应不同硬件和软件配置，提高方案的通用性和适应性。

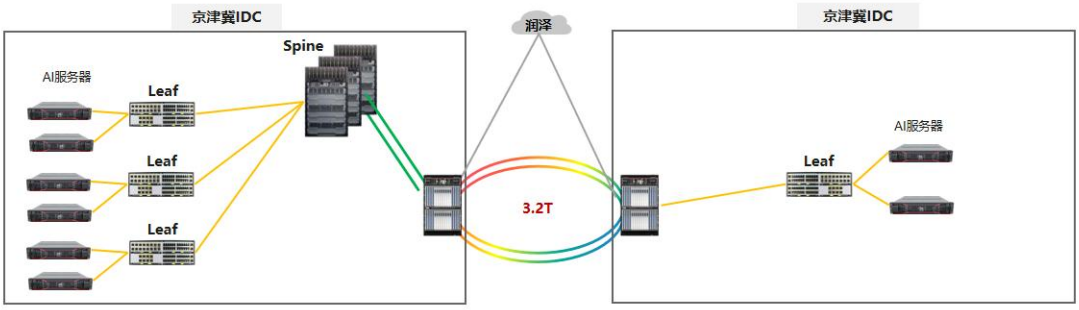


图 4-1 京津冀智算机房 80km/120km 绕行拉远验证组网

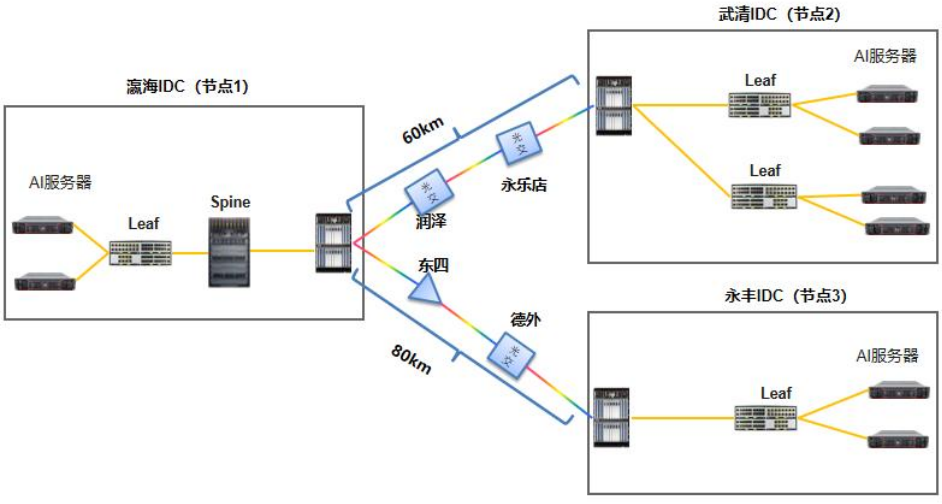


图 4-2 武清、瀛海、永丰三地 IDC 机房拉远验证组网

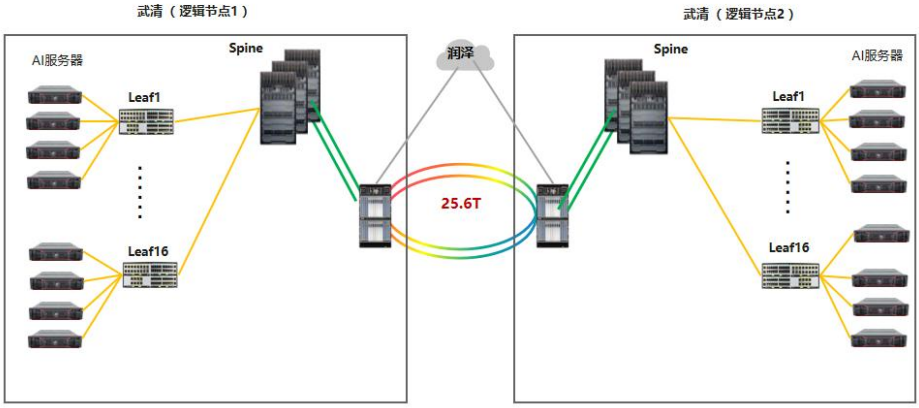


图 4-3 京津冀智算机房千卡 120km 绕行拉远验证组网

4.3 试验结论

项目组利用分布式智算中心无损网络方案整合 DC 机房资源，在全球首次解决了百公里长距跨机房大模型训练难题。训练效率方面，在不同组网拓扑下不同模型跨机房训练均可达同机房训练性能的 95%以上，证明分布式智算中心无损网络的可行性；网络稳定性方面，分布式智算中心无损网络可支持大模型一轮 5000 次迭代训练任务，均完成超 12 小时、约 80w 条样本数据的稳定性测试，具备支持大模型长期稳定训练的能力。分布式智算中心无损网络测试验证及相关创新研究将助力多方小规模智算中心并联成虚拟的大型智算中心节点，实现区域内智算中心协同计算模式，解决临时性的大规模算力需求，推动端网算协同创新，解决供给与需求区域发展不平衡问题，促进京津冀战略协同，快速推进智算中心建设，夯实新一代算力底座，为区域算力互联网的建设打下坚实基础。

5 总结和展望

面对新时代、新业态、新要求，中国电信积极践行“以网强算”的技术路线，打造面向智算业务的新型基础设施，以高性能智算网络作为提升集群算力性能的关键抓手，突破智能算力供给瓶颈。

本白皮书从智算业务的典型需求和特征、分布式智算中心无损网络方案、关键技术、典型实践四个方面开展了相关研究。未来，随着算力需求的持续增长，

分布式智算中心无损网络将进一步依托国家项目“多模态智联计算网络技术研究与验证”中的核心技术，在赋能智算基础设施方面发挥更加重要的作用，为经济社会发展注入新的动力。

附录 A：术语与缩略语

英文缩写	英文全称	中文全称
AI	Artificial Intelligence	人工智能
DCN	Data Center Network	数据中心网络
DNS	Domain Name Service	域名服务
DWDM	Dense Wavelength Division Multiplexing	密集波分复用
ECMP	Equal-Cost Multipath	等价多路径路由
FLOPS	Floating-point operations per second	每秒浮点运算次数
GPU	Graphics Processing Unit	图形处理器
HPN	High-Performance Network	高性能网络
IB	InfiniBand	“无限带宽”技术
IP	Internet Protocol	网际互连协议
LCOS	Liquid Crystal On Silicon	硅基液晶
LLM	Large Language Model	大语言模型
NTP	Network Time Protocol	网络时间协议
RDMA	Remote Direct Memory Access	远程直接数据存取
RoCE	RDMA over Converged Ethernet	融合以太网承载 RDMA
RSVP	Resource ReSerVation Protocol	资源预留协议
RTT	Round-Trip Time	往返时延
TCP	Transmission Control Protocol	传输控制协议
TPU	Tensor Processing Unit	张量处理器
OCH	Optical Channel	光信道
ODU	Optical channel Data Unit	光通道数据单元
OSNR	Optical Signal-to-noise Ratio	光信噪比
OTN	Optical Transmission Network	光传输网
WSON	Wavelength Switched Optical Network	波长交换光网络
WSS	Wavelength Selective Switch	波长选择开关

附录 B：参考文献

[1] 以网补算，构筑智算时代新底座 [EB/OL](2024-5-50)[2024-8-1]. https://xueqiu.com.

[2] Google. Gemini: A Family of Highly Capable Multimodal Models, 2024.

[3] Kun Qian, Yongqing Xi, Jiamin Cao et al. Alibaba HPN: A Data Center Network for Large Language Model Training, 2024.

[4] 百度智能云.智算中心网络架构白皮书[R/OL](2023-6).

[5] Microsoft. Empowering Azure Storage with RDMA, 2023.

扫码下载此电子书

中国电信股份有限公司研究院 I 2024.8