中国电信
CHINA TELECOM

基于光电协同的智算网络技术白皮书

中国电信股份有限公司研究院

2025年8月

版权声明

中国电信股份有限公司研究院
联系人：解云鹏
联系电话：010-50902166
邮箱：xieyp6@chinatelecom.cn

编写组

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广东分公司、中国电信股份有限公司上海分公司、中国电信股份有限公司北京分公司

参与编写单位：

华为技术有限公司
中兴通讯股份有限公司

主要编写人员（排序不分先后）：

中国电信股份有限公司:
傅志仁、雷波、唐静、李聪、解云鹏、张越、冀思伟、吴楠、马小婷、王飞飞、孙吉斌、赵倩颖、卫敏、邢文娟、王艺霏、潘永琛、曾涛、陈新豪、罗明军、王浩、方鸣、白洋、姚凌、王轶

华为技术有限公司：
陈文波、孙光辉、邵旭龙

中兴通讯股份有限公司：
冯志坚、谢大、李连华

高级顾问（排序不分先后）：
刘志军（中国电信股份有限公司广东分公司）
张婷（中国电信股份有限公司广东分公司）
张坚平（中国电信股份有限公司上海分公司）
叶平（中国电信股份有限公司北京分公司）

目录

前 言

2025 年 3 月，政府工作报告提出将持续推进“人工智能+”行动，支持大模型广泛应用。这意味国家将进一步加强顶层设计，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，弥补与国外在算力规模上的差距，夯实智算业务发展基础。

本白皮书围绕智算业务的典型需求和特征，对基于光电协同的智算网络相关方案、核心技术深入研究，并积极推动入算、算内和算间网络的现网验证。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光电协同智算网络。

1. 智算时代的业务需求分析

基于 Transformer 架构的通用大模型在近年迎来新的发展爆发期，AI 大模型竞争进入推理强化和应用拓展的进一步激烈竞争[1]。2025 年 4 月 Meta 发布的 Llama 4 Behemoth 总参数已高达 2 万，Llama 4 Scout 模型支持的 1000 万 token 的上下文窗口更是刷新了开源模型处理长文档、复杂对话和多轮推理任务能力的纪录[2]；OpenAI 已发布的 GPT-5 有 30 至 50 万亿参数量，较 GPT-4 实现性能上质的飞跃，万卡级集群成为训练标配，1.6T 光模块与全光交换技术将通信时延压缩至微秒级，而分布式训练架构的成熟则推动跨数据中心算力协同成为新常态。模型规模和复杂度的持续提升正重塑智算产业底层需求，亟需构建算网基础设施底座，优化算网服务系统，打造智算互联网新体系。

1.1. 智算业务的发展与特征

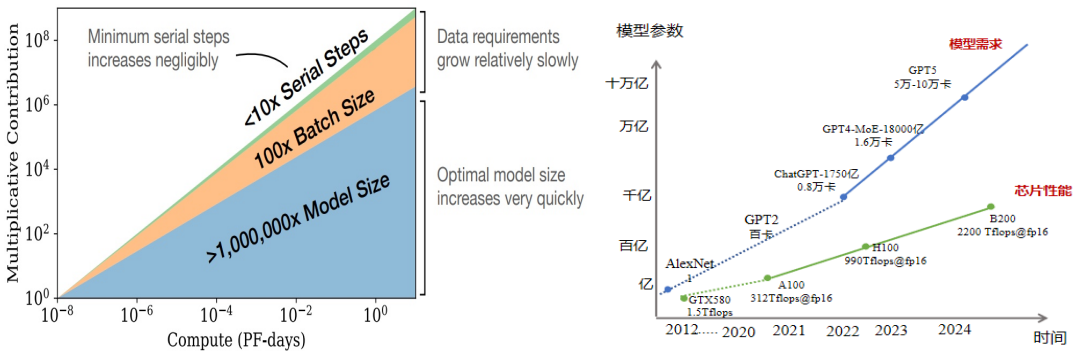


图 1-1 (a) AI 大模型扩展规律示意图 (b)模型规模与硬件性能发展趋势

智算算力需求目前呈现出前所未有的指数级增长态势。从规模上看，参数量突破是提升大模型性能的关键，模型规模的提升使得其语言理解、内容生成、逻辑推理等智能任务的处理能力显著跃升。图 1-1 展现了影响大模型性能的因素，可以看到模型性能随计算量、参数量、数据量指数增加而线性提高，大模型迭代过程中参数也逐年激增；从结构上看，人工智能的突破式发展推动了传统算力供给模式的系统性重塑，即由 CPU 为主的通用算力演变为“通智超”一体化供应的算力结构，成为塑造人工智能领域新模式、新业态的核心驱动力。

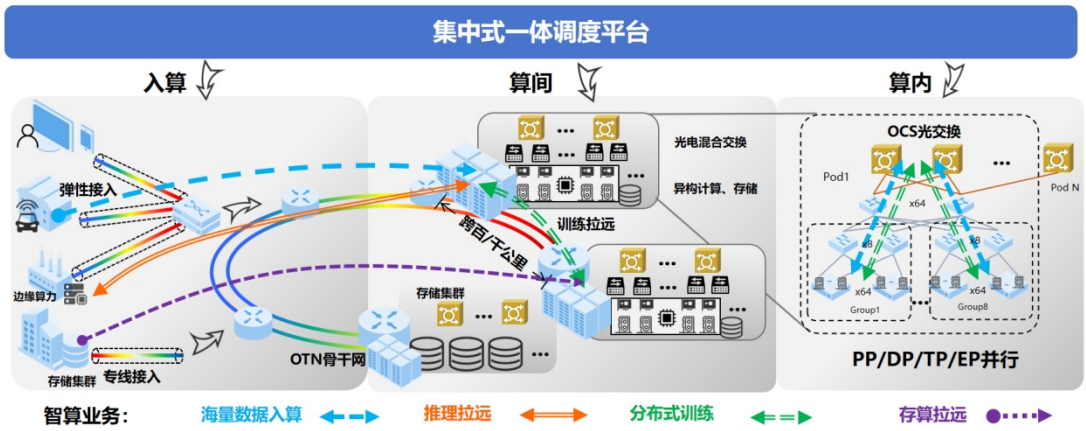


图 1-2 智算网络整体视图

作为智能算力的主要载体，智算网络支撑着人工智能相关新型业务演进中的技术进步和行业应用，智算业务规模的扩大更是对底层网络提出了更高的需求。如图 1-2 所示为基于大模型业务的智算网络整体视图，其以 AIDC 为核心，服务于数据入算、模型训练、推理等典型场景，提供一体化智算服务的新型信息基础设施，旨在支持大规模、高带宽、低时延和高可靠性的智算业务。

在数据入算场景中，海量数据需被上传至算力中心，作为后续模型训练与推理分析的基础数据支撑。网络覆盖从用户端至智算中心的全链路，由于数据量巨大、对网络指标（如时延）需求差异化明显，亟需大带宽、高弹性的入算网络支持；在模型训练场景中，主要对大规模数据基于复杂算法进行深度处理，使模型逐步掌握数据中的规律与特征，从而形成具备特定智能能力的模型，具备完成预测或决策任务的能力。此阶段需要海量算力，主要依赖单一大规模智算中心或多个智算中心间的高性能网络连接，覆盖区域和骨干网络；在模型推理场景中，主要面向实际业务应用场景利用已训练的模型对新输入数据进行实时分析与预测，实现智能化的决策支持或服务输出。其在网络方面需要支持用户到智算中心间的通信，同时覆盖多个推理池间的数据网络，以满足海量用户高并发、低时延的推理服务需求。

1.2. 智算业务对光电协同的需求

智算业务的蓬勃发展对光电协同智算网络提出了多层次的创新要求：

在入算场景中，面对海量数据的高弹性、高吞吐需求，需构建高速泛在的入算网络体系。依托超宽接入能力打造高速弹性智联架构，实现数据高速入算、带宽动态调整及端到端质量的全链路保障；

在算内网络层面，需聚焦光电深度融合，创新提出基于 OCS（Optical Circuit Switch，光路交换机）的新型组网方案。通过光电协同技术部署超大规模集群网络，全面提升集群的性能、横向扩展性与组网灵活性；

在算间网络领域，需建设超高速全光互联体系，通过新型光纤技术试点应用，突破光传输技术在算间通信中的性能瓶颈，同时前瞻性探索更大容量的新一代全光网技术，为算间协同奠定技术基础。

光电协同网络旨在深度融合光通信的高速传输优势与电处理的灵活管控能力，构建面向智算业务全生命周期的一体化网络支撑体系。其核心功能聚焦于为大规模智算场景提供全流程保障，覆盖数据接入、计算处理到服务输出的完整链路，持续输出高效、可控、可靠的网络能力。该网络依托“光电算用”一体化调度机制，实现从数据接入、节点间协同计算到节点内部资源调度的端到端贯通，最终构建起以业务需求为驱动、以高效计算支撑为核心目标的智算网络基础设施底座。

1.3. 光电协同国内外业界现状

在光电协同智算网络领域，已有多家企业建立面向智算业务的新型网络基础设施，目前有多项落地案例与推进计划，体现了技术创新与应用需求的深度融合。

国外企业如英伟达引入 OCS 技术构建了可重构光互联网络，基于光开关实现 AI 集群的动态调整与灵活扩展，实现集群动态调整与故障快速恢复；谷歌应用 OCS 技术优化拓扑结构，基于 3D-Torus 架构构建 TPU 集群，其中 TPU v4 已实现集群可用性从 8%至 75%的跃升[3]，最新的 TPU v7 集群延续了此架构[4]，基于六千卡规模智算组网，完成了对 PaLM 大模型的训练；日本电信公司 NTT 基于 APN 实现跨多数据中心 LLM 远程训练，验证 APN 技术在跨地域算力调度中的可行性。国内如华为推出数据中心全光交叉解决方案，通过光电混合架构与动态拓扑调度能力，同时基于全光交换机，将端口可靠性提升至 20%以上。

综合来看，目前基于光电协同的智算解决方案已展现出了显著优势：通过突破传统网络架构瓶颈，能够有效降低数据传输时延并提升网络响应速度，同时为智算中心内部组网提供了更灵活的拓扑设计与更高效率的连接能力。然而，业界在智算网络领域缺乏系统化的端到端解决方案，以实现从数据接入、算力调度到多节点协同处理的全流程高效支撑，确保智算业务在未来大模型时代下的持续扩展与高效运行。

2. 光电协同智算网络总体解决方案

智算网络总体解决方案通过光电协同全栈创新，构建入算网络、算内网络和算间网络的全方位、端到端智算网络，旨在为智算业务提供精准业务感知、灵活调度策略、高效数据传输及可靠安全保障等关键能力。

2.1. 方案设计原则

光电协同智算网络总体解决方案在设计时，应遵循超高通量、超高可靠、灵活扩展、平滑演进的设计原则：

（1）超高通量原则

为满足智算业务海量数据处理及高速运算需求，智算网络架构需采用超高容量通信链路及高效网络交换设备，支持极高吞吐量与极低延迟，确保数据在各层级网络间快速、流畅传输，避免网络瓶颈影响智算系统整体性能。

（2）超高可靠原则

智算业务对数据准确性与连续性要求极高，任何网络故障都可能引发严重后果。因此，智算网络架构需构建完善的可靠性机制，涵盖 IP+光的冗余链路设计、设备备份以及故障自动切换等功能。通过这些措施，确保部分网络组件出现故障时，整个智算网络仍能正常运行，保障业务连续性与数据完整性。

（3）灵活扩展原则

鉴于智算业务规模持续增长，智算网络架构应具备良好可扩展性，采用模块化的组件设计，在不影响整体网络运行的情况下，支持网络拓扑结构和网络层级的灵活扩展，灵活增加或替换网络节点与设备，以适应未来业务发展的不确定性。

（4）平滑演进原则

随着业务发展及技术迭代演进，在升级带宽、扩展规模时，传统方式往往需更换大量设备和线缆，成本高且实施难度大。智算网络架构设计充分考量光通信在大带宽、长传输、低功耗等方面的优势，引入多芯光纤、光交换机等，在无需大规模更换基础设施的情况下，实现网络的平滑演进，保护前期投资。

2.2. 光电协同智算网络总体架构

如图 2-1，光电协同的智算网络的整体架构由“四层三域”构成，横向划分为四层，分别是网络设施层、网络能力层、网络调度层以及业务应用层；纵向划分为三域，包括入算网络域、算间网络域和算内网络域。

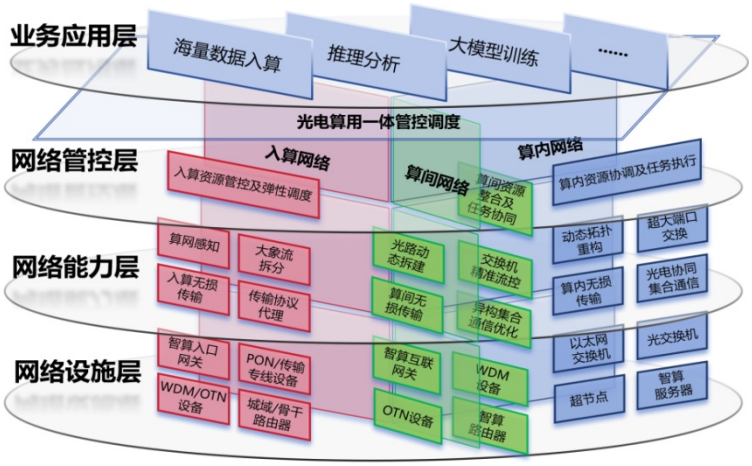


图 2-1 基于光电协同的智算网络总体架构

其中四层包括：

网络设施层：作为数据传输和计算的承载平台，基于智算网关、智算路由器、智算服务器、光电交换机、光接入设备、光传输设备等构建“入算-算间-算内” 的端到端传输路径，为网络能力层、网络管控层、业务应用层提供**物理基础支撑**。

网络能力层：依托算网感知、大象流拆分、无损传输、协议代理、光路动态拆建、精准流控、集合通信优化、拓扑动态重构、超大端口光交换等技术，保障数据“快速、准确、可靠”传输，为网络管控层、业务应用层提供**高效通路保障**。

网络管控层：智算网络的大脑，实时监测网络设施层、网络能力层状态，结合业务应用层需求，基于光电一体协同管控能力，对入算网络设施、算间网络设

施、算内网络设施进行管控调度，实现业务需求与底层资源高效匹配，是连接业务应用与底层资源的**资源协调中枢**。

业务应用层：基于用户场景与需求，调用网络管控层分配的资源开展业务处理。以模型训练为例，业务应用层发送任务至网络管控层，触发资源分配与数据传输流程，驱动整个智算网络运转，是智算网络运转的**需求驱动核心**。

三域包括：

入算网络域：处于智算网络的边缘层，主要负责将外部数据源（如各类传感器、互联网数据等）接入到智算系统中。入算网络域的网络设施通常包含智算网关、PON、传输专线、WDM、OTN、城域路由器、骨干路由器等设备，将分散的数据整合后，上传至算内网络或算间网络，为后续的计算处理提供数据基础。

算内网络域：位于智算节点内部，是智算服务器/超节点之间进行数据交互与通信的核心网络。除了智算服务器和超节点，算内网络域的网络设施还包括以太网交换机和光交换机，负责在单个智算节点内部，协调各类计算任务的执行，保障计算资源的高效利用与数据的快速处理，是智算系统内部运算的关键支撑网络。

算间网络域：连接多个智算节点，实现智算集群之间的通信与资源共享。算间网络域的网络设施通常包含智算互联网关、WDM、OTN、智算路由器等设备，支持分布式计算任务的协同执行，将各个智算节点的计算能力进行整合，实现大规模智算任务的并行处理，提升整个智算网络的计算能力与业务处理范围。

各层和各域之间相互协作、相互依赖，形成一个有机的整体。纵向维度上，入算网络域、算内网络域、算间网络域通过网络设施层的传输链路贯穿全系统，成为各层间数据流转的物理脉络；横向维度上，网络设施层的硬件资源、网络能力层的技术支撑、网络管控层的智能决策、业务应用层的需求指令，通过功能衔接形成价值传递链。这种纵横交织的架构中，网络管控层作为中枢枢纽，既衔接业务应用层的动态需求，又调度底层网络域的资源供给，推动全系统按需协同，最终实现从智算数据接入到业务落地的高效运转。

3. 入算网络关键技术

大模型训练、推理、数据挖掘等新型智算业务迅猛发展，算力的规模化、异构化和集约化趋势日益显著。智算业务呈现出数据规模大、实时性强、资源调度复杂等特点，直接推动了入算网络架构与能力的演进。

3.1. 入算网络技术总览

入算网络是智算网络的重要组成部分，负责将分布在不同地域、不同接入层的海量数据高效汇聚至智算中心，为大规模人工智能训练、推理及跨区域算力协同提供高带宽、低时延、稳定可靠的传输通道。

为满足海量数据入算的性能与稳定性需求，入算网络必须具备高带宽、低时延、零丢包、动态弹性、安全可靠等能力。围绕这些需求，入算网络的关键技术可以概括为以下四类：

- **算网感知技术：**通过对应用、算力和网络状态的实时感知，入算网络能够识别不同智算任务的带宽、时延及可靠性等需求，根据计算资源分布和网络实时信息，进行最优路径选择和任务调度。
- **弹性带宽技术：**面向智算业务流量的突发性和阶段性特征，弹性带宽技术通过光层大颗粒带宽调度与 IP 层细颗粒流量调优相结合，实现秒级乃至亚秒级的带宽动态扩缩容。结合网络切片能力，不同租户、不同业务可获得定制化的带宽保障，提升网络资源利用效率并降低运营成本。
- **无损传输技术：**针对分布式训练等高性能计算场景，入算网络需实现端到端零丢包与低抖动传输。通过基于 RDMA（Remote Direct Memory Access，远程直接数据存取）、RoCE（RDMA over Converged Ethernet，基于融合以太网的远程直接内存访问）等高性能协议的精准流控、拥塞感知与反馈机制，结合硬件级流特征感知与智能识别技术，消除网络瓶颈对任务效率的影响。同时，针对非 RDMA 环境，可通过智算网关进行协议转换，保障异构网络间的高效互通与无损传输。
- **安全可靠技术：**面向跨地域、多租户和多业务的入算场景，网络必须具备从物理层到应用层的多重安全防护能力。通过“量子+OTN”技术实现抗量子计

算的物理层全业务加密，保障高安全场景下的低时延、零带宽损耗传输。结合多速率混合 ROADM（Reconfigurable Optical Add-Drop Multiplexer，可重构光分插复用器）组网技术，灵活承载 100Gb/s 至 800Gb/s 多速率业务并提升网络恢复成功率至 98%，为高带宽业务提供稳定可靠的光传输保障。

通过上述四大关键技术的协同，入算网络能够在接入、城域与骨干各层次实现对智算应用的全流程保障，成为支撑未来智算业务发展的核心支柱。

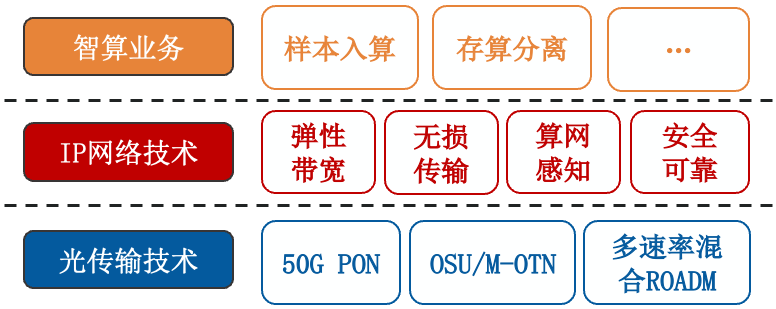


图 3-1 入算网络关键技术架构

3.2. 算网感知技术

智算业务覆盖的应用场景呈现出高度多样化，不同应用对于算力和网络的需求存在差异，并且计算模式、数据分布和通信模式呈现动态变化的特征。因此，作为承载智算业务的入算网络需要具备全面且精准的感知能力，能够实时感知应用需求、算力资源状况及网络状态，满足智算应用复杂多变的计算需求，保障计算任务的连续性和高效性。

（1）应用感知

应用感知旨在识别和理解不同智算业务的运行特征与需求模式，包括模型类型、计算强度、数据传输规模、实时性要求等，从而为算力与网络资源的精准匹配提供依据。

应用感知技术可以通过应用感知网络协议和部署智算网关来实现。常见如 APN6 技术，利用 IPv6 协议扩展的字段携带应用标识和需求信息，包括计算资源需求、网络带宽、时延需求等。边缘侧的智算网关捕获应用流量的需求，与光网算用平台进行协同，实现业务资源之间的精准匹配和高效调度。

（2）网络感知

在入算网络中，网络感知旨在精准掌握链路的时延、带宽、丢包及拥塞状态，从而为智算任务调度、路径选择和网络质量保障提供实时依据。

在时延感知方面，广泛采用随流检测 iFIT、带内网络遥测 INT 等技术，实现毫秒级至微秒级的单向、双向时延与抖动探测；在带宽感知方面，则结合带内带宽遥测、基于速率探测的可用带宽测量、端口统计采集等精准获取链路可用带宽和利用率。

（3）算力感知

算力感知旨在关注算力节点（GPU/TPU、加速器、存储节点等）资源状态与可用性，确保智算任务计算时分配到最适合的节点，提高整体算力效率。

算力感知技术实现有多种方式，包括利用集中式的算力管控平台，采用算力节点上报方式获取算力的实时状态；或者算力资源侧部署智算网关或节点代理 Agent，通过特定的采集协议实现算力状态的实时同步。

3.3. 弹性带宽技术

在智算业务场景中，数据流量呈现出高度的突发性与不均衡性特征，传统的固定带宽配置模式无法兼顾资源利用率与业务性能保障，容易导致高峰期链路拥塞、低谷期带宽浪费。

弹性带宽技术聚焦于构建低成本、灵活可弹、高速的数据传输网络。弹性带宽技术包括以光网络为底座的物理层弹性和网络层基于网络切片的弹性，通过光电弹性能力的协同，为智算业务提供端到端的带宽稳定性和可预测性。

（1）光层弹性带宽技术

光网络作为承载层核心，凭借其大容量、低时延和高可靠性特性，为带宽弹性调整提供了坚实的物理基础。光网络的带宽调整主要依托于可调波分复用（WDM）与可编程光传送技术，通过灵活配置波长资源、调制格式及传输速率，实现动态带宽扩展。

接入侧采用 50G PON 光接入技术，基于单波长 TDM-PON 架构，上行 TDMA/下行 TDM，通过 ONU 和 OLT 动态缓存技术方案，可以为用户数据入算提供万兆接入能力，实现大突发场景中无损保障。引入计算周期和下发周期动态调整，实现低时延能力保障。

OSU（Optical Service Unit，光业务单元）是面向城域优化的光传送技术，通过灵活映射、低时延封装，为智算时代提供低时延、高可靠、灵活带宽的综合业务承载。在低时延优化层面，突破传统 OTN 五层冗余封装，创新采用“OTUCn-ODU4-OSU”三层极简架构，优化业务封装时延和单站穿通时延，满足 AI 推理的毫秒级响应需求。在带宽灵活性层面，支持带宽颗粒动态调整，通过并行调速技术实现秒级无损带宽变更。这种“低时延+全灵活”的技术特性，为智算时代的全光运力网络提供低时延、高可靠、灵活带宽配置的综合业务承载。

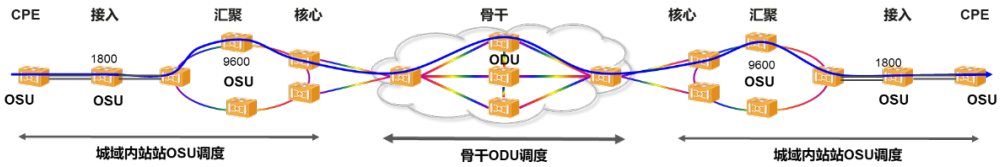


图 3-7 M-OTN/OSU 端到端调度方案图

（2）网络层弹性带宽技术

在智算场景中，不同用户、不同任务对带宽的保障等级和时延敏感度存在显著差异。

基于 FlexE（Flexible Ethernet，灵活以太网），入算网络可在同一物理基础设施上构建多个逻辑隔离的虚拟网络，每个切片拥有独立的带宽、时延与可靠性保障策略，从而精准匹配应用带宽需求。同时，结合 SRv6（Segment Routing over IPv6，技术进行切片路径的精细化编排），可确保高优先级智算任务在网络拥塞或故障场景下依然获得带宽优先保障，而低优先级业务则可在剩余资源上弹性运行。

通过 DPI（Deep Packet Inspection，深度包检测）、流量特征分析以及基于 AI 的流量预测算法，IP 层可实时识别智算任务流量模式，提前预留或调整资源，确保大规模分布式训练、模型推理及跨地域数据同步等场景的网络性能。

3.4. 无损传输技术

为满足智算业务对高吞吐、低时延和零丢包的极致传输需求，构建面向流级精细感知与跨层协同的无损传输体系。该体系通过大象流拆分技术实现硬件级特征采集、智能识别与跨层负载分担，缓解流量集中造成的拥塞瓶颈；通过协议转换技术打通传统以太网与 RDMA 网络，实现异构环境下的高性能无缝传输；并

依托层级化负载均衡技术，结合跨层映射、精准拥塞控制与 SDN 自动化编排，实现端到端零丢包和资源最优调度，为智算业务提供稳定、高效、可扩展的传输保障。

3.4.1 大象流拆分技术

智算业务的流量呈现“大象流占比高、流特征低熵化、多流同步化”的典型特征，构建“硬件级流特征感知+智能算法识别+跨层负载拆分”协同技术机制。硬件级引擎集成模块采集流特征并标记同步行为，智能算法通过深度分析特征实现报文区分与大象流识别，跨层负载分担采用 QP（Queue Pair，队列对）+五元组协同拆分技术，满足智算业务无损传输需求。

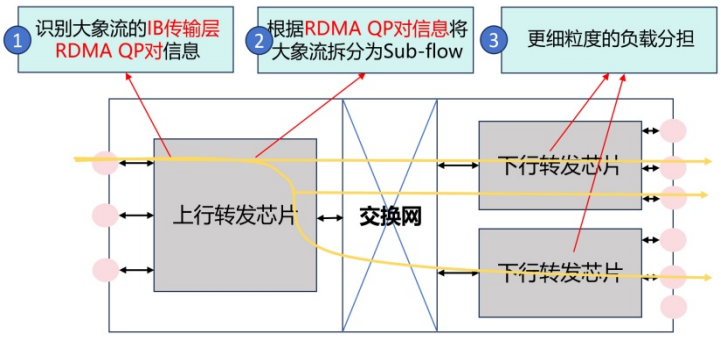


图 3-3 大象流识别与拆分架构图

3.4.2 协议转换技术

为保障智算业务在异构网络环境中的高性能传输需求，入算网络需要具备流级无损能力，数据传输采用 RDMA、RoCE 等面向高吞吐与低时延的无损传输协议，通过引入精细化流控和拥塞检测等机制，可以形成端网协同的无损传输能力，有效避免传输路径中的数据丢包和时延抖动问题。

考虑到传统数据中心仍大量采用 TCP/IP 协议栈，缺乏原生 RDMA 支持，入算网络引入智算网关作为协议转换枢纽，实现从传统以太网到 RDMA 网络的无缝衔接。智算网关具备基于 DPDK（Data Plane Development Kit，数据平面开发套件）的高性能数据面与 RoCEv2 协议栈的深度融合能力，可在接收到传统 TCP 流时，进行流级解包、缓存聚合与流识别，并以 RDMA 方式重构数据传输路径，从而绕过内核协议栈，降低延迟，提高带宽利用率。智算网关可与调度系统协同，

基于算力流优先级及业务特征动态配置流控策略，配合网络侧无损机制，确保传统数据中心业务在进入 RDMA 网络后同样具备流级无损保障。

3.4.3 层级化负载均衡技术

智算场景下单条转发路径需承载大量并发流，为突破传统负载分担中应用层流特征不可感知导致的流间资源竞争问题，构建跨层映射、多维度均衡及 SDN（Software-Defined Networking，软件定义网络）自动化编排协同技术机制。通过 RDMA QP 对与 IPv6 跨层映射技术，结合 YANG 模型实现流级信息标准化上送与策略自动化下发，突破流级调度瓶颈。

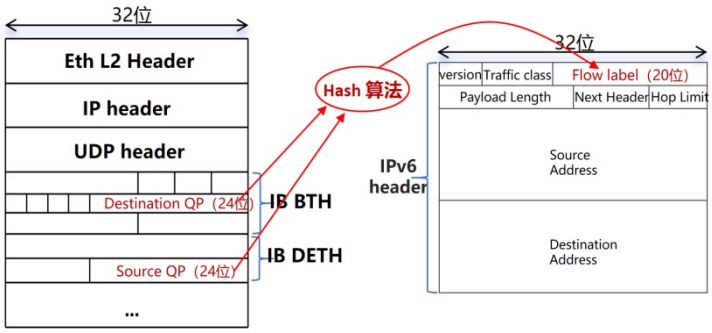


图 3-4 流级负载均衡图

智算业务对网络拥塞高度敏感，1e-4 丢包率就会导致有效吞吐下降超 50%，因此需具备细粒度精准拥塞控制能力以缩小影响域。通过基于 ICMPv6（Internet Control Message Protocol version 6，互联网控制消息协议第 6 版）报文扩展传递租户级反压信息实现精准流控保障零丢包，并将其与数据中心内 PFC（Priority-based Flow Control，基于优先级的流控制）技术基于优先级映射互通，构建跨广域端到端精准流控能力。

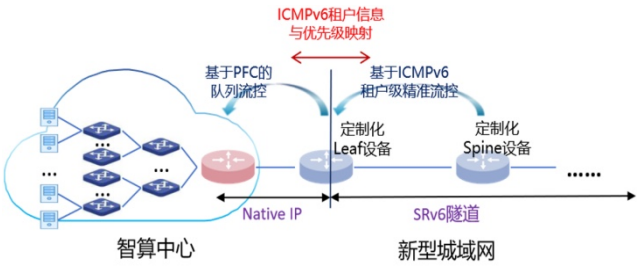


图 3-5 租户级负载分担图

3.5. 安全可靠技术

面向智算时代的高安全与高可靠传输需求，提出量子+OTN 一体化方案，通过引入独立 QKD（Quantum Key Distribution，量子密钥分发）网络与 CSP（Cryptography Service Platform，密码服务平台）集中密钥管理，实现抗量子计算的物理层全业务加密，兼顾微秒级时延、零带宽损耗与灵活带宽适配能力。同时，基于多速率混合 ROADM 组网技术，实现 100Gb/s 至 800Gb/s 多速率业务的灵活接入与智能调度，结合光层故障精准定位与恢复机制，将业务恢复成功率提升至 98%，为高带宽业务提供稳定、韧性且可持续演进的光传输能力。

3.5.1 量子+OTN 技术

针对党政军、金融、数据中心互联等高安全场景，传统加密方案存在加密层级有限、业务适配范围窄、传输时延高、带宽损耗大等瓶颈，且难以对抗量子计算对加密体系的威胁。“量子+OTN”技术通过依托独立 QKD 网络生成量子密钥，由 CSP 密码服务平台集中管理“量子密钥池”；OTN 侧 QCPE（Quantum Cryptography Customer Premise Equipment，量子加密客户终端设备）设备实时在线申请业务密钥，支持周期更新，并通过离线充注完成设备认证鉴权，保障交互安全，实现量子密钥与 OTN 业务协同调度。OTN 网络与 QKD 网络整合，安全层面突破抗量子计算壁垒，构建物理层全业务加密能力，提升 OTN 专线安全韧性；性能层面实现微秒级时延、无带宽损耗及灵活带宽适配，支撑高实时性、大带宽业务场景。

3.5.2 多速率混合 ROADM 组网技术

多速率混合 ROADM（Reconfigurable Optical Add-Drop Multiplexer，可重构光分插复用器）组网技术聚焦智算时代骨干光纤网多速率业务混合承载及高可靠运维难题。ROADM 设备及网络架构支持 100Gb/s、400Gb/s、800Gb/s 等多速率光波长信号的混合接入、调度与传输，通过光层的灵活重构能力，实现不同速率业务的动态适配、高效承载与智能运维。100Gb/s 网络依托冗余 OTU（Optical Transport Unit，光传送单元）端口提供检测光源，400Gb/s 及以上速率网络通过

填充光构建检测通道，二者协同软件调度技术实现恢复路由性能检测与故障精准定位。该技术有效增强网络可靠性，持续提升波道恢复成功率，在网络资源适配场景下，可支撑单点光缆故障业务平均恢复成功率 98%，为多速率混合业务的稳定运行筑牢技术根基。

4. 算内网络关键技术

随着人工智能技术的快速发展，大模型训练参数量呈指数级增长，单数据中心内的算力规模正从百卡级向万卡级甚至十万卡级演进。这一趋势对算内网络提出了更高要求，需要其具备超大规模组网、无损高吞吐传输和智能容错等核心能力。在此背景下，构建面向未来的超大规模智算中心已成为产业发展的迫切需求。

4.1. 算内网络技术总览

算内网络是指数据中心内部连接 GPU/XPU 等算力节点的互联网络，承担分布式训练中梯度同步、参数聚合等通信密集型任务。其核心功能是为大规模智算节点间的数据交换、分布式计算任务、高速参数同步等提供高带宽、低延时、高可靠的网络，保障集群算力线性扩展。

算内网络是智算中心的基座，直接影响算力效率，是算力效率提升的关键瓶颈之一。算内网络主要为胖树架构，AI 调度平台调度算卡，网络控制管理网络设备。通过光电协同互联架构，为智算参数面提供端到端的低时延、高带宽、低功耗、可编程的网络。其中，算内网络的关键技术主要包括：

- **光电混合互联架构：**融合电交换机与 OCS，实现灵活拓扑重头与带宽按需供给。
- **超大端口光交换技术：**基于 OCS 的高容量、低功耗交换能力，实现高密光互联，超低功耗光交换。
- **光电协同控制技术：**亲和调度算法，光交换与电交换的动态协同机制，采用动态路由策略和多路径负载分担，实现算力和网络资源协调，算力充分利用。
- **集合通信库算法优化：**基于 OCS 架构的通信算法创新，实现带宽充分利用。

4.2. 光电混合互联架构

在算内网络中，结合电交换机和光交换机形成新型光电混合架构，能有效兼顾网络的灵活性和高带宽。该技术解决的主要问题是单一网络架构在面对超大规模算卡集群时难以平衡性能与灵活性的困境。当前，单纯依赖电交换机的架构，在大规模集群下易受限于电信号传输的带宽瓶颈和较高的时延，且随着算卡数量激增，其功耗和成本也会大幅上升；而纯光交换的架构虽能提供高带宽和低损耗，但在动态连接调整上不够灵活，难以适应算卡间频繁变化的通信需求。

然而，实现这些目标面临着多重挑战：**速率**方面，受限于电交换机的芯片速率，面向 AI 流量普遍出现带宽“瓶颈”，且因负载分担通信关系复杂导致网络利用率偏低；**路由协同**方面，光交换为端口级交换，其静态路径配置与电交换的动态路由协议需通过上层管控平面进行协同；**能效平衡**方面，传统交换机网络端口密度与功耗呈指数增长，光交换的低功耗特性与电交换的灵活性需要权衡；**时延限制**方面，电交换机多跳转发时延增加，影响了通信效率。

光电混合架构的创新在于融合两类交换机优势，通过统一调度实现协同。技术原理体现四大优势：

- 超大规模：光交换机以高端口密度和大带宽构建高速链路，支撑十万卡级以上集群，电交换机承担小范围高密度接入；
- 超低时延：光交换机直传光信号减少电转换时延，电交换机通过协议与缓存优化降低小范围交互延迟，协同降低端到端时延；
- 灵活扩展：电交换机动态响应临时通信需求，光交换机依负载调整预设光路，提升扩展与升级灵活性；
- 超高可靠：采用冗余链路设计，主链路故障自动切换备用链路，结合备份设备与毫秒级快速路由切换算法，确保通信持续稳定。

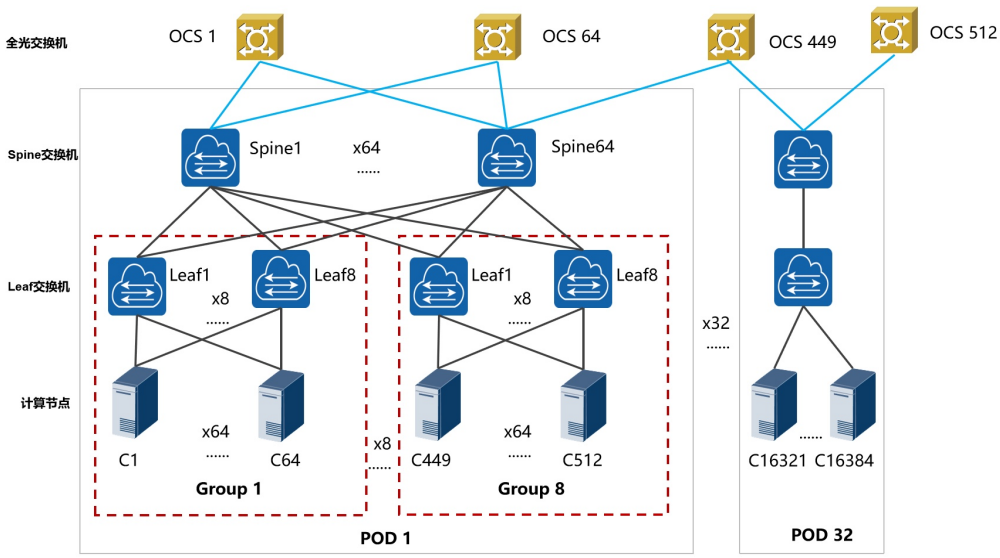
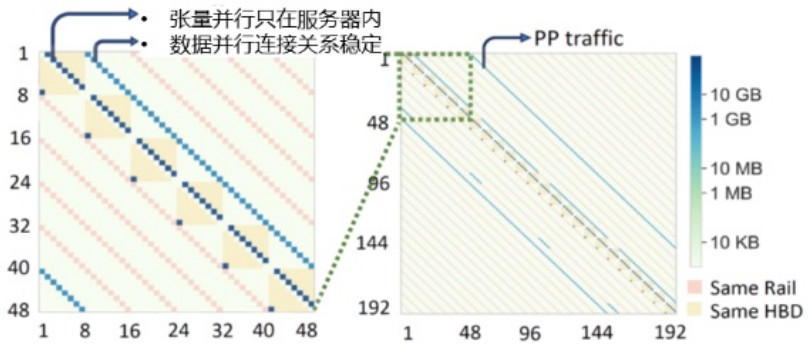


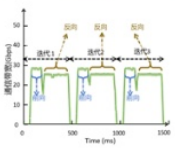
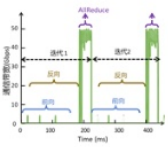
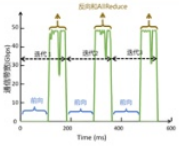
图 4-1 光电混合互联架构示意图

该架构具体原理是，光电混合互联架构根据大模型流量规律，根据其流量可预测的特性，设计该架构。大模型采用分布式并行训练方式，主要包括张量并行（TP）、流水并行（PP）、混合专家并行（MOE）、数据并行（DP），由此产生的 AI 训练流量具有明显规律。



来源： MIT & Meta《How to build low-cost networks for large language models》

图 4-2 AI 训练流量关系矩阵

	张量并行（TP）	流水并行（PP）	数据并行（DP）	混合专家并行（MOE）
流量模型				流量模型与TP类似
通信操作	All reduce	Send-receive	All reduce	AllToAll
通信范围	服务器内，一般为8卡	组内或组间同号卡	组内或组间同号卡	组内同号卡

来源：MIT《Network-Aware Job Scheduling in Machine Learning Clusters》

图 4-3 并行训练特点

从 AI 训练流量关系矩阵可以看出，通信关系具有局部性，仅有部分算卡之间存在通信，大部分算卡之间无通信，因此只需采用 OCS 连接需要通信的卡或 POD。

依据并行训练特点，AI 训练流量具有高度的周期性且可预测，其中张量并行和数据并行为 All Reduce 通信，可通过算法调整具体通信顺序；流水并行为点到点通信，通信关系单一；专家并行为 ALL2ALL 通信，需要通过算法控制通信范围。由于 AI 训练流量具有局部性、规律性强、可预测、可规划的特点，能够提前通过分析模型确定通信关系和通信流量，因此可在模型启动训练或部署时配置 OCS，按需调整网络拓扑适配模型。

基于上述技术原理。OCS 作为光路交换设备，可用于承载 POD 间经汇聚形成的稳定大象流，实现 POD 间互联与跨 POD 大流调度，构建点到点高带宽通道，可减少转发跳数、降低时延，并削减传统核心层电交换机数量及设备维护等资本和运营支出。架构采用分层混合组网：电交换机在接入侧处理细粒度实时变化流量，OCS 提供跨机架/跨 POD 专用高速通道，实现带宽、灵活性与低功耗的统一。同时，基于 Spine - Leaf 架构，结合多链路、多设备设计及软件定义的链路健康检测与自动切换机制，保障网络高可靠性。

4.3. 超大端口光交换技术

光电混合架构的核心内容是在架构中引入了光交换层，可以为数据中心网络创建任意的逻辑拓扑结构，同时打破互联带宽的限制。汇聚层通过光纤连接到光交换机，通过配置每个光交换机来连接输入和输出光纤的排列组合，进而实现不同逻辑拓扑。基于 SDN 进行动态管理和实时运维，打破现有数据中心架构形态，实现光层可重构。光交换的技术目前主要有 3D MEMS（Micro-Electro-Mechanical Systems，微机电系统）技术和液晶光交换技术等

（1）3D MEMS

3D MEMS 光交换技术，即三维微机电系统技术，是基于传统的 MEMS 技术基础上发展而来，融合了半导体的核心在于利用微机电系统，利用微型机械结构实现光学信号路由。控制 MEMS 反射镜阵列，将来自源端口的入射光束重定向到目标端口，完成了对于传统 EPS（Electronic Packet Switch，电路分组交换机）的替代，并实现了低延时、传输速率透明、低功耗三方面的性能提升。

3D MEMS 光交换技术通过三维可控小型反射镜动态引导光信号，在光纤间灵活建立或断开连接，实现高效光信号管理。其核心优势主要有三点：

- 微镜独立操作支持多信号高效路由，无接触设计减少磨损、提升可靠性；
- 可构建大规模交换矩阵，具备高扩展性；
- 低光学损耗优化系统效率，且能快速响应网络变化实现光路实时重配置。

该技术适用于数据中心互连、电信骨干网/城域网、高性能计算平台等场景，既提升现有网络性能，也为未来高速大带宽通信需求奠定基础。

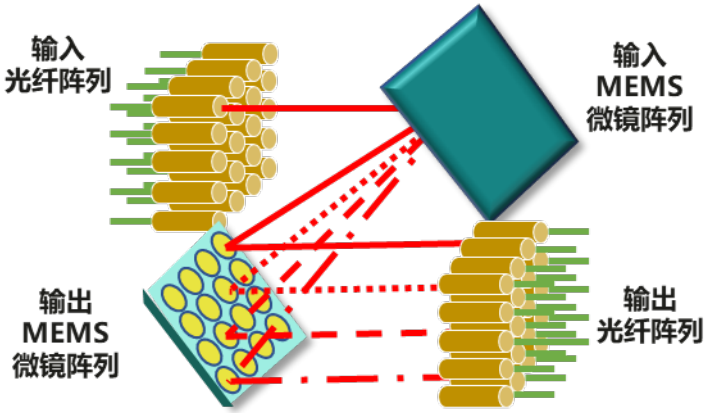


图 4-4 3D MEM 原理示意图

（2）液晶光交换技术

液晶光交换技术，特别是基于 LCOS（Liquid Crystal on Silicon，硅基液晶）的技术，是一种利用液晶材料和半导体制造工艺实现的先进光学信号处理方法。该技术通过在硅片上集成一层液晶层，并施加电信号来控制液晶分子的方向，从而改变光线路径或调制光信号，特别适用于需要高精度和灵活性的应用场景。LCOS 采用反射式架构，光线穿过液晶层到达下方的反射面后，依据液晶分子的状态被调制并反射回来，这种设计不仅提供了更高的分辨率和对比度，还允许更高的开口率和更精细的像素控制。在光通信网络中，LCOS 技术用于构建动态可重构的光交换系统，如波长选择开关（WSS），这些设备能够在多个输入输出光纤之间动态分配不同的波长通道，极大地提高了网络的灵活性和资源利用率。

其优势在于无机械运动部件，避免了传统机械式光开关常见的磨损问题，延长了使用寿命，同时低插入损耗确保高质量的光信号传输，然而该技术响应速度慢于 3D MEMS，对网络状态的实时感知较弱，不适合在快速切换的应用中使用，且成本相对较高。因此，3D MEMS 光交换技术成为目前全光网络的核心基础，

当前商用 3D MEMS 光交换机已支持 320×320 端口，未来通过硅光集成和微镜阵列微型化，有望实现 1000×1000 以上规模，满足超大规模智算中心需求。

4.4. 光电协同控制技术

在大规模异构算力网络中，设计光电协同的互联和路由方案是提升集群带宽利用率、降低延迟的关键。电交换机为包交换，支持路由协议，交换灵活度高，然而能耗高、链路拓展受限，存在负载突发拥塞问题：OCS 为端口交换（光纤交换）适合大带宽、流量规律场景，需要由控制平面配置光交换的链接；光电交换设备互为补充，动态分工，最大化实现“弹性网络”，为此，设计拓扑感知的智能亲和调度算法和动态光电路由协同策略。

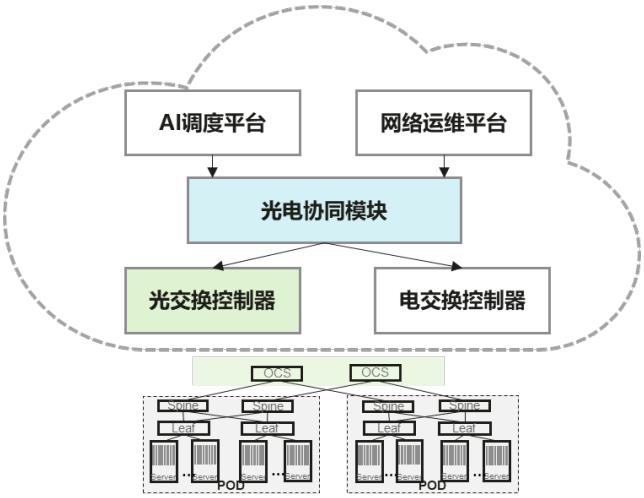


图 4-5 算光电协同管控架构

AI 调度平台分析 AI 模型确定流量关系矩阵，将流量关系矩阵发送给光电协同模块；光电协同模块计算根据各并行需要的拓扑，配置 OCS 在 POD 间组成最佳拓扑。此外，光电协同模块根据 POD 内的服务器分配情况和 OCS 连接后的网络拓扑计算网络路由，防止出现拥塞，按照路由配置电交换机进行通信。

4.4.1. 拓扑感知的智能亲和调度算法

利用 AI 流量局部性、规律性强、可预测可规划的规律，AI 训练通信流数少，但带宽大、规律性强的特点使用光交换直连，对 AI 任务节点做“带宽亲和性”分组，调整网络拓扑适配 AI 任务，减少密集通信的跳数，降低通信时延，提高通信效率。

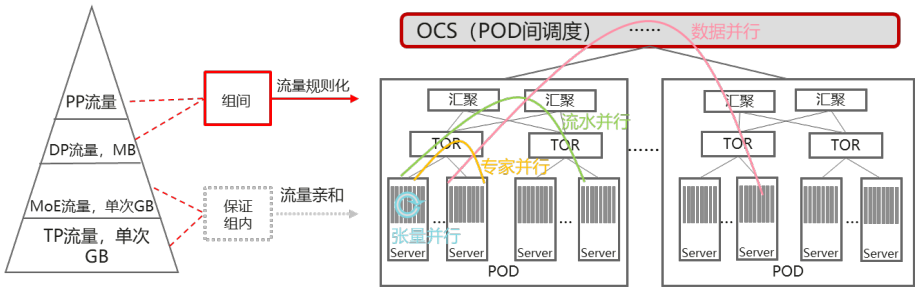


图 4-6 参数面流量及亲和调度策略

所有流量优先部署在组内减少通信层次；张量并行一般部署在服务器内，少量会部署在服务器间。需要跨组部署时，专家并行流量大时延有严格要求，PP 流量小但通信次数多，DP 流量比 PP 大，通信次数少；因此按照组内部署的优先级为专家并行>PP 并行>DP 并行的亲和调度策略在 AI 集群中部署模型，提升算力利用率。

4.4.2. 光电路由协同策略

由于电交换机网络普遍采用 ECMP 进行带宽负载分担，因此进行链路 Hash 时经常出现冲突，如下图。

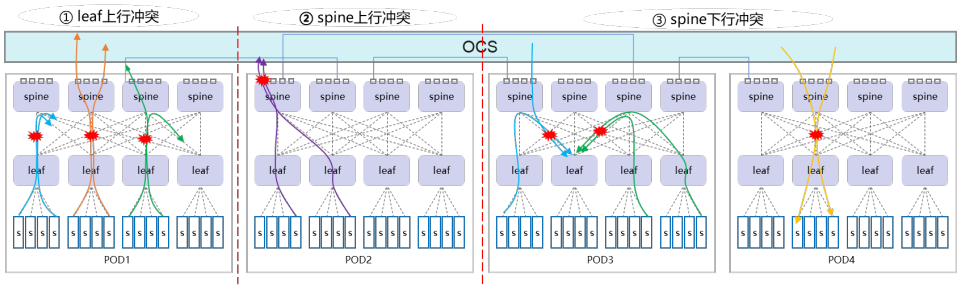


图 4-7 流量冲突示意图

在 Leaf 节点上行、Spine 节点上行和 Spine 节点下行都可能出现 Hash 后多条流抢占一个端口的情况，Hash 冲突导致网络带宽无法重复利用，时延增大，严重时会导致丢包，严重影响通信性能。因此需要管控平面根据调整后的网络拓扑，自顶向下编排电交换机的路由，实现组内和组间按需通信、互不干扰，最大化利用通信带宽，避免流量冲突。

根据 POD 间流量情况配置 POD 间的链路，使网络拓扑适配模型。同时根据网络拓扑和各并行训练的通信关系设计互相通信的智算卡的通信路径，并配置路由，防止进行负载分担时，多条流同时抢占部分链路，而另一部分链路控制，导

致链路带宽不能充分利用，影响通信效率。根据实时流量和端口健康动态分配连接方式；如遇链路拥塞或异常，自动进行路由切换。

4.5. 集合通信库算法优化

在大规模分布式 AI 集群中，集合通信作为带宽消耗核心，其与底层 OCS 架构的适配及精细化调度，是提升网络性能的关键突破口。当前集合通信库基于电交换机胖树架构设计，依赖逐包交换能力，因缺乏对光电混合架构的适配，易导致带宽浪费或拥塞。

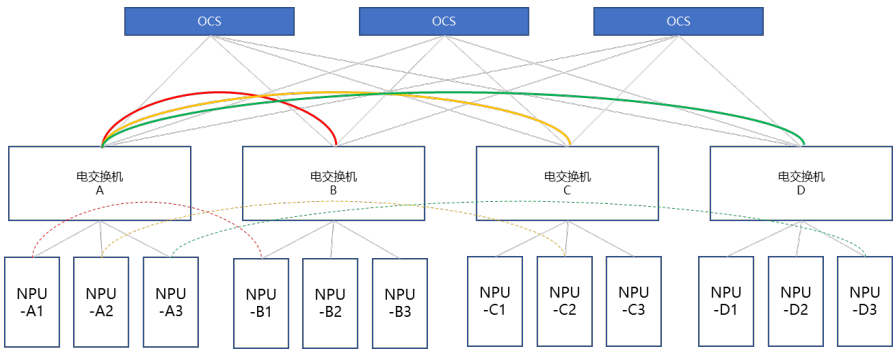


图 4-8 基于 OCS 的集合通信适配原理图

OCS 具有链路配置固定性，链路一旦建立仅支持特定对象通信，无法实现带宽共享。以 12 张智算卡组成的集群为例，4 台交换机各下挂 3 张卡，每台交换机通过 3 个端口经光纤连接 OCS，网络无收敛。模型部署需全量卡通信时，存在同交换机下 3 张卡间及跨交换机对应序号卡间的通信需求。交换机 A 通过 3 个端口分别与 B、C、D 交换机建立链路，若 A 下挂 3 张卡同时与 B 下挂 3 张卡通信，会出现单链路抢占而其他链路空闲的带宽不均问题，导致通信性能下降。

优化核心在于通过集合通信库算法调整适配 OCS 特性：首个周期 A1 与 B1 通信时，A2 与 C2、A3 与 D3 通信；下一周期 A1 与 C1 通信时，A2 与 D2、A3 与 B3 通信，依此类推实现动态通信配对。通过调整集合通信库的通信顺序，可充分利用集群网络带宽，同时发挥 OCS 低时延优势，有效提升通信效率。

5. 算间网络关键技术

随着模型参数规模从亿级提升到万亿级别，训练模型所需的算力资源也呈指数级增长，智算中心集群规模向十万卡级甚至百万卡级加速演进。然而，单一智算中心因其机房空间、散热、电力供应等实际因素无法满足大模型训练的计算需求，跨多个数据中心的分布式训练成为重要的发展方向。由于静态算内网络拓扑无法适配分布式训练、推理场景，因此需要通过 IP 层和光传输层技术的协同优化来构建高性能算间网络，实现跨广域的分布式训练、推理。

5.1. 算间网络技术总览

算间网络是指连接多个智算中心内计算节点、存储资源的高性能互联网络，其核心目标是打破物理地域限制，实现算力、数据、存储资源的高效协同与池化调度。它以低时延、高吞吐、无损传输为核心特征，依托全局负载均衡、精准流量控制、智能拥塞控制等 IP 管控技术和大带宽、低时延、高可靠的光传输技术，构建“算力-数据-网络”深度融合的互联体系，支撑分布式 AI 训练、微调、推理等智算场景。算间网络在智算网络中承担“跨域互联枢纽”角色，总体视图如下：

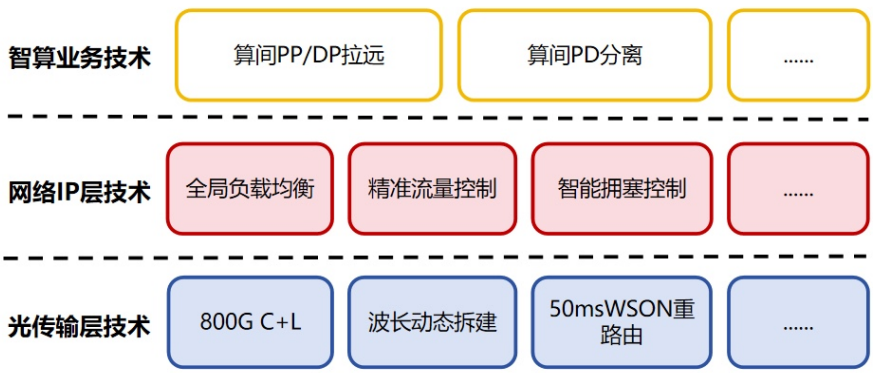


图 5-1 技术逻辑视图

5.2. IP 层管控技术

5.2.1. 全局负载均衡技术

分布式训练、推理场景中，由于带宽、时延的非对称性和异步性，传统负载均衡技术难以实现流量的均衡负载分担。全局负载均衡技术通过对 DC 内和跨 DC 流量的统一规划，确保所有路径流量均衡无冲突，避免拥塞丢包，具体流程如图 5-2 所示。首先网络设备从服务器获取 GPU 间通信矩阵关系并上报控制器；随后控制器结合算内和算间的训练流、推理流及组网拓扑进行集中式算路，为每条流量选择合适路径；最后控制器将路径信息下发到各网络设备，网络设备根据计算结果对算内和算间的路径进行动态调节。

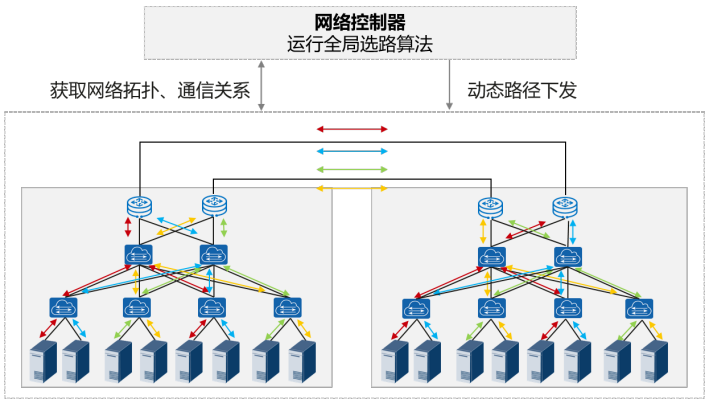


图 5-2　全局负载均衡过程

5.2.2. 智能拥塞控制技术

基于 ECN（Explicit Congestion Notification，显式拥塞通知）的拥塞控制机制在单数据中心内拥塞控制效果明显，被广泛采用。但在分布式训练、推理场景下，长距链路发生拥塞后，过长的拥塞反馈路径会导致源端服务器流量不能及时降速，加剧设备拥塞程度。

I-ECN（Intelligent Explicit Congestion Notification，智能拥塞控制）技术将长距链路拥塞转移至网络第一跳设备，通过缩短拥塞反馈路径并控制源端速率来缓解拥塞，具体流程如图 5-3 所示。首先网络设备实时监测网络状态，判断链路是否发生拥塞；若链路发生拥塞，则拥塞点向离源端服务器最近的网络设备（设备 A）发送通告报文。随后设备 A 根据报文信息判断链路拥塞程度，并计算需要

向源端服务器发送的 CNP 报文或者其他流控协议报文的数量；最后源端服务器调整对应流量发送速率，实现合理控速。

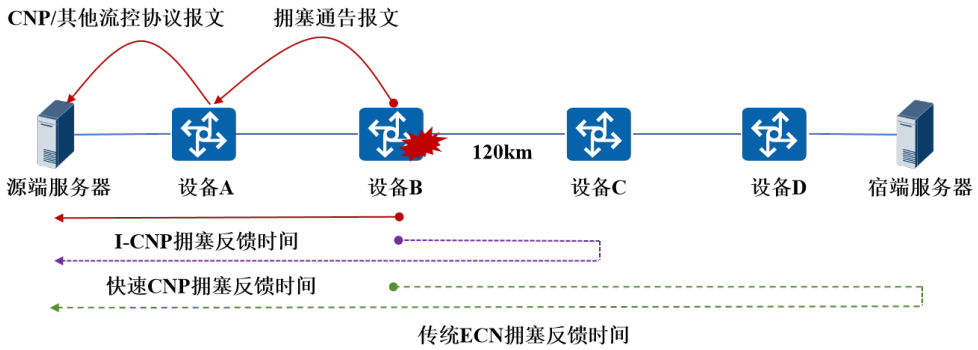


图 5-3　智能拥塞控制过程

5.2.3. 精准流量控制技术

分布式训练、推理场景下，网络环境更加复杂多变，传统 PFC（Priority-based Flow Control，基于优先级的流量控制）机制会引发 PFC 头阻、反压风暴和死锁问题，进而导致拥塞在多租户间扩散。

精准流量控制以 IP 数据报文的五元组作为流识别粒度，实现了对每条数据流的独立监控与动态调节，降低了拥塞扩散的风险，具体流程如图 5-4 所示。首先为每条队列设定反压阈值，并实时感知网络的拥塞状况；随后当队列占用缓存超出预设的反压阈值，设备会迅速生成流控反压报文，并通过反向路径通知上游设备暂停该队列的数据传输。最后当该队列缓存降至反压阈值以下时，拥塞设备解除拥塞状态，并停止向上游设备发送流控反压报文，上游设备重新发包，实现基于队列的无损弹性传输能力。

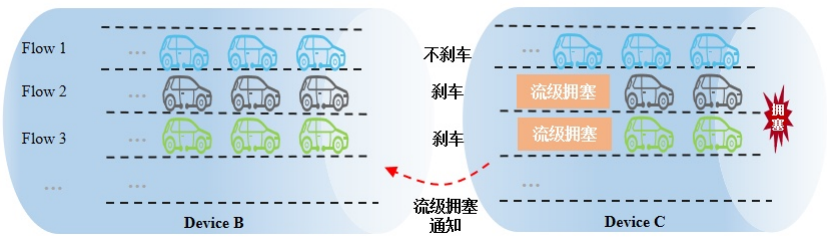


图 5-4 精准流控技术

5.3. 光传输技术

5.3.1. 800G C+L 传输

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，并部署在智算 DCI 百公里级互联场景中，在满足百 T 级大带宽需求的同时，降低了智算互联的成本。

除了提高单波长速率外，扩展传输波段也是提高互联带宽的有效途径。通过将频谱资源扩展到 C+L 波段，并配合中短距单波长 800G 技术，可以实现最大 96Tbit/s 的单纤带宽，进一步满足智算中心之间的海量数据传输需求。

图 5-5 C+L 波段提供更大容量

5.3.2. 波长级动态拆建技术

从实际业务来看，智算资源一般采用分时复用的方式租给不同的用户，因此需要在任意两个智算中心之间根据空闲 GPU 数实现带宽弹性互联，即光传输网络需具备波长级的动态拆建能力。

如图 5-6 所示，网络根据用户的需求匹配合适的 GPU 数量，并依据应用场景（如距离、带宽等），从业务侧驱动建立不同方向的波长级连接。结合当前的网络拓扑与网络资源的利用情况，波长动态拆建技术需具备如下能力：

（1）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（Optical Channel，光通道）预开通路径；

（2）自动生成所配置的业务参数，包括客户侧到 OCH，OCH 到光纤的中心波长频率、线路侧单波长速率、端口配置与多层路由映射等；

（3）基于当前拓扑的结构与实际情况，将开通的 OCH 自动调优至最佳状态。

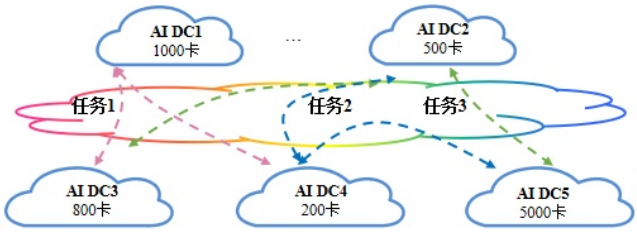


图 5-6 带宽分时复用的业务场景

5.3.3. 50ms WSON 重路由技术

传统 WSON（Wavelength Switched Optical Network，波长交换光网络）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，而秒级的断纤时间至少会损失 30%的效率，甚至会导致训练中断。因此，需要提升 WSON 的重路由能力，将重路由时间控制在 50ms 内，实现确定性的光层恢复能力，确保集群整体的计算效率。50ms WSON 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（Wavelength Selective Switch，波长选择开关）快速切波技术。

- **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。
- **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。
- **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。
- **WSS 快速切波技术：**通过使用全新的快速液晶材料以及 LCOS（Liquid Crystal on Silicon，硅基液晶）技术实现 ms 级的波长交叉切换能力。

6. 光网算用一体化调度平台

光网算用一体化调度平台将深度融合光网络、IP 网络、算力资源与大模型应用，构建“光-网-算-用”四维一体的智能调度体系。该平台以光网络、IP 网络、算

力资源为底层基础设施，通过实时感知、智能分析、全局优化和动态协同，实现网络资源、计算资源与应用需求的精准匹配，最终构建高效、弹性、可靠的广域智算网络生态，支撑大模型训练/推理等多样化 AI 业务的差异化需求。

6.1. 光网算用一体化调度架构

光网算用一体化调度架构采用“适配层-感知层-控制层-服务层-全链仿真”五层设计，通过分层解耦实现“感知-决策-执行”闭环，支撑入算、算内、算间调度需求，充分发挥光可靠传输、网无损承载、算高效利用优势，满足多样化大模型差异化应用需求。其中，光网算用一体化调度架构主要包括：

适配层：通过协议解析（如 SDN/RDMA）消除异构系统壁垒，解决多源异构系统的兼容性问题，实现光、IP、算力资源、以及智能应用等的统一接入，支撑上层业务对网络能力的灵活调用；

感知层：对资源与网络状态的实时感知的信息进行量化与智能分析，构建“算力-网络”双域状态的数字镜像，为控制层提供精准决策依据；

控制层：作为一体化调度系统的智能调度中枢，包括任务调度、网络调度、资源调度、SaaS 部署、自动化调度等功能，实现任务与资源调度；

服务层：提供任务管理、算力租赁等标准化接口。包括提供任务全生命周期管理的任务中心、负责对外输出算力 API 能力的算力中心、提供网络连接服务的网络中心、封装大模型应用的智能应用中心、以及评估与优化业务运行效果的综合评估中心。

全链仿真层：为了降低大规模集群试错成本并提升部署成功率，本调度架构增加了全链仿真层，通过数字孪生预演任务执行过程，优化资源策略。

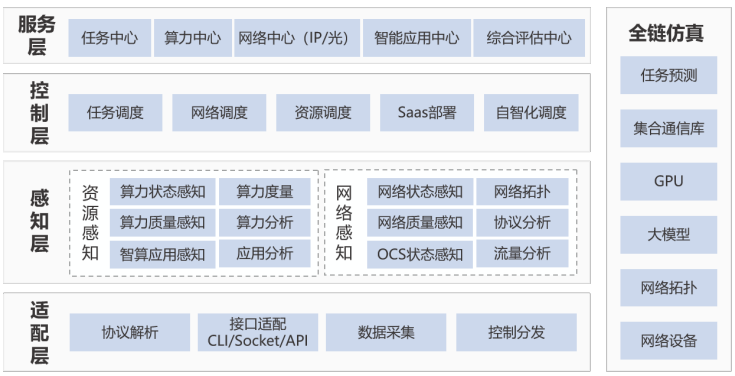


图 6-1 光网算用一体化架构示意图

6.2. 大模型任务调度优化

区别于传统互联网公用网络的流量无序性，分布式智算网络凭借相对封闭的网络环境和充足的任务模型部署信息，依托能力层实时采集的多维度资源数据，构建包含计算性能、存储成本、通信开销的算网资源度量标准模型，形成具备自适应能力的动态算网资源描述语言。

基于感知层对网络拥塞状态、节点负载水位、链路传输时延的实时监测能力，调度平台可结合 Transformer 架构中 PP 流水线并行、DP 数据并行、TP 张量并行等分片策略的资源需求特征，对大模型训练所需的分布式计算资源和节点间传输资源进行多周期预测，精准评估不同并行策略组合下的训练效率与资源消耗。通过构建“资源需求-网络状态-策略匹配”的智能决策引擎，平台可在任务提交阶段快速完成算网资源的全局寻优，动态匹配最优的并行分解方案，并通过动态调整数据传输路径、优化跨设备卸载策略，实现异构资源的高效协同。该机制不仅突破了传统调度中资源匹配的粗放式瓶颈，更通过将存储成本、计算性能、通信开销纳入统一评估模型，显著提升了 GPU 集群在长序列处理、多模态融合等复杂场景下的资源利用率，为大模型训练提供了从任务分解到算网协同的全链路智能化支撑。

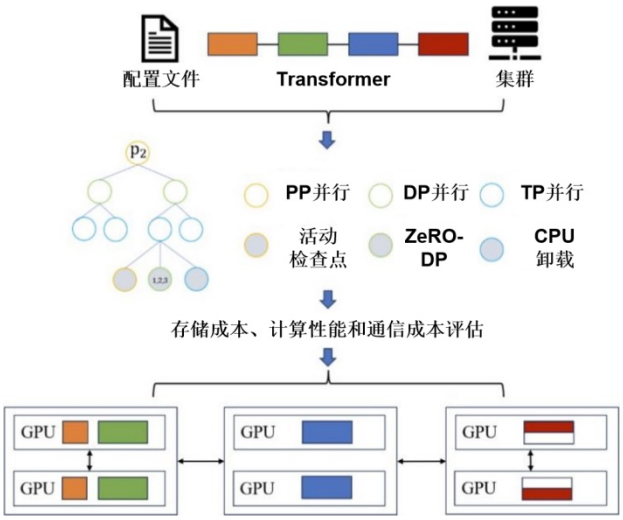


图 6-2 大模型任务分解示意图

6.3. 算网资源协同调度技术

分布式大模型训练任务的拆解方式和子任务的卸载位置决定了智算节点之间的流量模型，流量模型与网络带宽的不匹配将会导致训练时间长、收敛速度慢等问题，同时为避免频谱碎片、负载不平衡等问题导致训练任务失败，迫切需要进行智算节点的算力资源与光层频谱资源、IP 层带宽资源和智算中心算力能力的协同调度。光网算用一体化调度平台结合训练数据并行和模型并行机制，通过感知智算节点的算力资源以及链路资源，制定表示算力资源和网络资源占用分布情况的指标，设计适应分布式大模型的资源调度算法，为拆分后的子模型选择合适的智算节点以及为目标智算节点间的数据传输确定路由与资源分配方案。

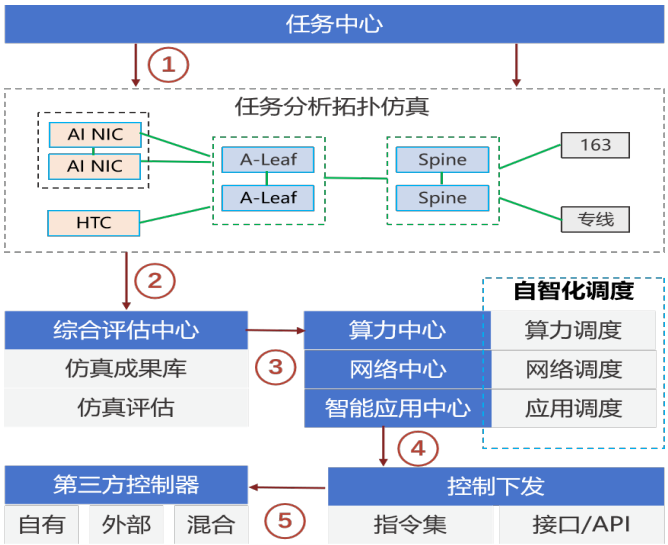


图 6-3 算网协同调度流程

6.4. 一体化管控技术

光网算用一体化调度平台需要考虑智算多维资源的高效管理、跨域智算任务分发和跨域流量调度，从而实现广域智算任务的分配和光、网、算、用资源调度的协同。

一体化管控包括智算网络域管理和资源管理的跨域智算网络管控，以及跨域网络管控实体和一体化调度平台及智算网络间的信息和控制信令及交互规范的定义和设计，并根据异构地址转换、翻译机制思想进行异构智算网络地址/标识空间的映射和管理，从而打通异构智算网络间依赖异构地址的域标识、资源标识

空间，支持广域智算网络光、网、算、用资源协同管控。其中域管理包括域加入、域发现、域退出机制等部分，资源管理包括资源上报、发现、资源同步和资源协商机制等部分。

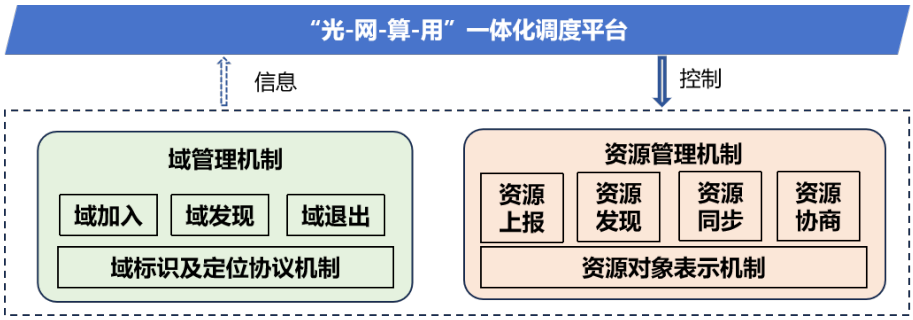


图 6-4 广域智算互联管控

6.5. 全链仿真技术

随着 AI 模型快速迭代发展，智算网络通信模型日益复杂，且随着集群规模向着万卡/十万卡/百万卡方向发展，无法搭建 1:1 物理环境验证。全链仿真模块将面向广域智算互联场景，构建高精度虚拟化网络与计算模型，致力于模拟大模型训练/推理任务中的通信时延、资源竞争与能耗特征，以指导广域智算互联组网、算法编排优化、网络资源配置决策与，低成本、快速对网络进行评估与优化，支撑方案创新和决策。

全链仿真包括作业仿真、网络仿真、虚拟算力资源、以及硬件集群信息、真实数据等部分。具体而言：

1）服务层将实际业务拆解为可仿真的需求、算力资源实际参数、网络拓扑、传输参数等，输入给全链仿真，模拟作业执行的全链流程；同时接收全链仿真输出的任务、网络、算力预测结果，辅助优化真实任务编排。

2）综合评估中心根据收集全链仿真产生的多维度数据（任务、算力、网络等仿真结果），进行综合分析，输出体系整体运行的评估报告（如效率、瓶颈诊断），同时将评估标准、目标反馈给仿真，校准仿真方向。

3）控制层将依据全链仿真结果制定各层调度策略，动态调整网络参数，并将调度指令同步给仿真进行策略有效性验证。

4）感知层可为全链仿真提供最基础、最实时的数据支撑，确保仿真模型中的资源状态与真实一致。

5）适配层负责解决全链仿真与底层能力层的数据、控制交互的兼容性问题。

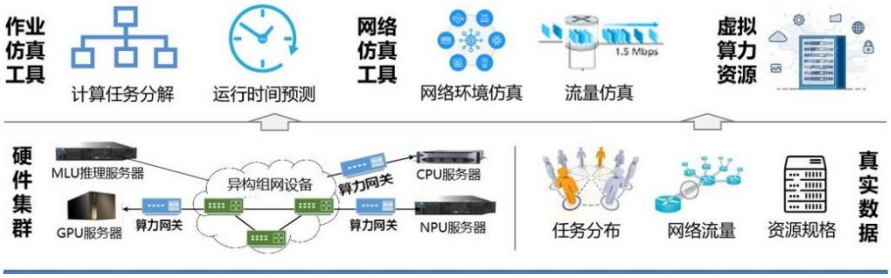


图 6-5 全链仿真思路图

7. 典型实践

在国家积极推动数字经济发展与新型基础设施建设的大背景下，《关于促进数据产业高质量发展的指导意见》明确提出到 2029 年数据产业规模年均复合增长率超 15%，国家发展改革委也着重强调要加快构建全国一体化算力网。人工智能的迅猛发展，使得各行业对算力需求呈井喷式增长，单体数据中心规模不断扩张，分布式集群部署趋势渐显，数据流量激增。

为响应国家政策，突破算力瓶颈，中国电信紧扣“光电协同赋能智算”这一核心目标，依托光电协同智算网络解决方案，在入算、算内、算间三大环节开展技术攻关，构建兼具超高带宽、超低时延、极致能效的智算网络体系，为全国一体化算力基础设施建设提供可复制的实践经验。

7.1. 入算试验

7.1.1 试验目标

存算分离是随着 AI 技术发展出现的新兴业务场景，企业将私有数据部署在本地，租用运营商智算中心算力卡进行存算分离拉远训练，采用基础大模型叠加私有行业数据进行二次训练和模型微调，要求网络具备弹性带宽、低时延、低成本、无损传输能力。

7.1.2 试验概述

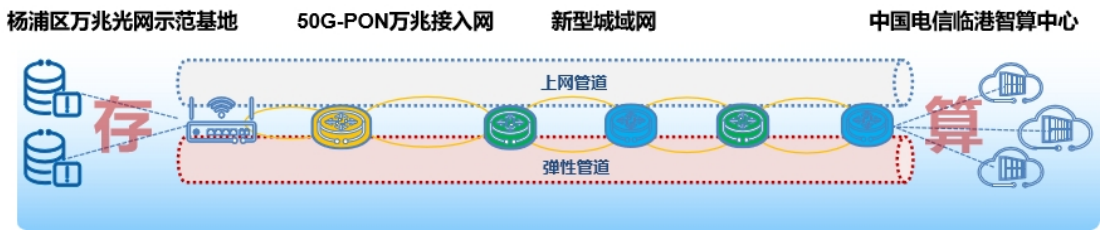


图 7-1 现网试验示意图

本次试验中，训练采用 Deepseek2-lite-16B 模型和存算拉远方式，基于 50G-PON 万兆接入网络，结合 RDMA 智算网关，打通了从用户到临港智算中心的 120 公里端到端网络链路，并以弹性智联网络方案架构为底座，提供 1G～25G 带宽按需可弹和端到端保障等技术特性。通过智能板实现了应用级业务质量实时可视，验证万兆光网高效承载训练过程中 20G 以上的突发峰值流量、及毫秒级入算的网络能力，确保用户到智算中心之间的广域高速无损传输。

7.1.3 试验结论

图 7-2 现网试验数据结果

现网试验数据表明：1）通过灵活加载业务控制模块，实现用户一键自主申请，分钟级完成业务开通；2）大容量样本数据上行带宽可超 10G，满足样本数据实时传输需求；3）样本入算业务传输时延从 30ms 降至 6ms。

7.2. 算内试验

7.2.1 试验目标

针对传统组网在超高速、大容量场景下面临的带宽、时延和灵活性瓶颈，基于 OCS 的新型光电协同组网架构凭借其大带宽、低时延和动态重构能力，显著提升了网络性能、扩展性与灵活性，为超大规模智算集群提供了高效可靠的组网解决方案。

本试验面向大规模智算中心，对比验证基于 OCS 的新型光电协同网架构与传统 CLOS 等主流拓扑在大模型训练场景下的性能优势，重点评估其在算卡性能、集合通信及模型训练等维度的有效性。

7.2.2 试验概述

本次试验依托电信实际网络环境，采用 LLAMA2-70B 大模型，针对 OCS 组网方案与纯电方案开展对比测试。试验核心内容包含以下三个内容：

1）光电混合功能测试：主要包含算光协同拓扑调整测试。

2）光电混合性能测试：主要包含算卡点对点性能测试、集合通信性能测试、加载稠密模型训练性能测试、加载稀疏模型训练性能测试。

3）纯电组网性能测试： 主要包含算卡点对点性能测试、集合通信性能测试、加载稠密模型训练性能测试、加载稀疏模型训练性能测试。

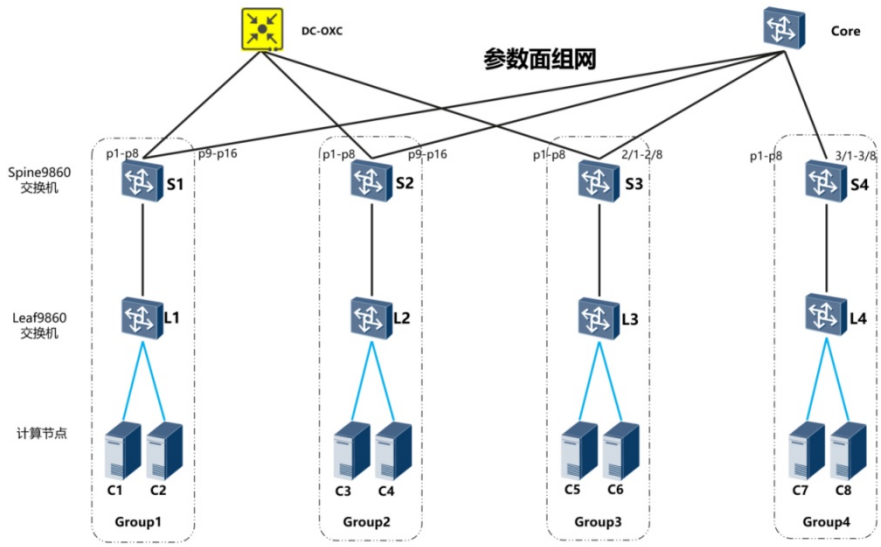


图 7-3 测试组网拓扑信息

7.2.3 试验结论

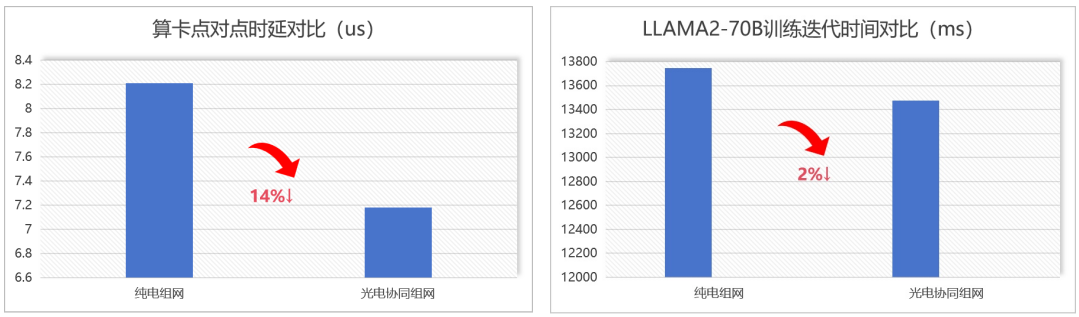


图 7-4 光电协同组网与纯电组网试验对比图

本次部署试验结果表明，基于 OCS 的光电协同组网方案相比纯电组网方案在算网融合核心性能指标上实现一定提升：在算卡点对点性能测试中，得益于 OCS 光层直连的低损耗特性，算卡点对点时延较纯电组网降低了 14%以上；模型训练性能方面，光电协同组网的迭代训练时间缩短 2%，且随着集群规模扩大（2048 卡及以上），性能优势呈现进一步放大趋势。上述结果验证了 OCS 技术在释放智算集群算力、优化通信效率方面的核心价值，为超大规模智算中心的无损互联提供了可行路径。

7.3. 算间试验

7.3.1 试验目标

为解决单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大模型分布式训练试验，验证跨数据中心合池训练的可行性，以提升区域内智算资源的供给效率。

此外，针对企业私有算力资源有限，难以满足海量用户并发的推理服务需求问题，通过多节点云、边协同，为用户提供无处不在的低时延推理服务。本试验面向分布式推理网络，验证 PD 分离架构下，广域 RDMA 无损等网络技术的可行性，支撑未来区域化智算推理集群的商用。

7.3.2 试验概述

1. 跨 DC 长距分布式训练

中国电信自 2024 年以来，多次在现网开展跨数据中心大模型分布式训练试验探索，并完成业界首个千亿参数大模型分布式联合训练试商用。基于大带宽、低时延和高可靠的光传输网络，以及全局负载均衡、长距无损流控等技术，使得集群计算指标接近本地计算场景。

项目组先后开展了五期大模型分布式训练试验。一期在京津冀智算机房进行两点 64 卡、80km/120km 绕行拉远试验，验证拉远距离对大模型训练性能的影响。二期在武清、瀛海、永丰三机房开展 64 卡、百公里大模型分布式训练，验证算间互联技术在多智算中心分布式训练场景下的效果。在前期百卡、百公里拉远验证基础上，三期在京津冀智算机房开展两点千卡、120km 拉远试验，探索长距链路带宽收敛情况下（带宽收敛比 4:1）模型训练的性能。四期、五期基于息壤一站式智算服务平台，分别开展千卡 120km/500km 拉远、带宽收敛比 16:1/32:1 下的广域分布式智算中心互联试商用，在互联距离、带宽收敛比及模型参数规模上均实现关键突破。此外，模拟了多种试验中可能出现的故障情况，以验证算间互联网络在面对线路路障、服务器端口故障及其他异常情况时的韧性和恢复能力。

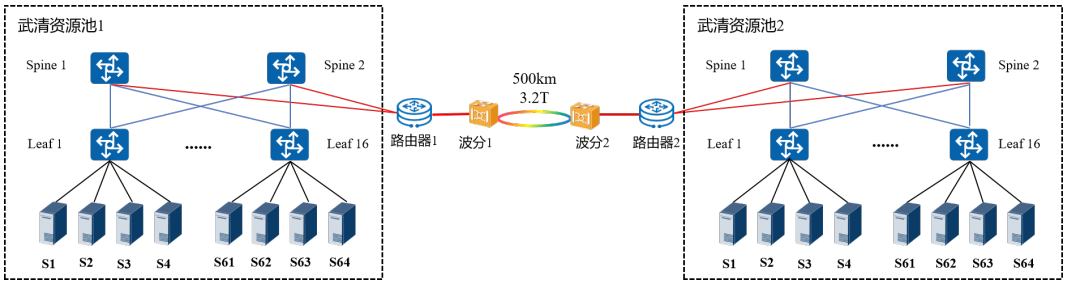


图 7-5 京津冀智算机房千卡 500km 绕行拉远验证组网

2. 跨 DC 分布式推理

2025 年 6 月，项目组基于广东 800G 波分设备完成业界首例长距跨 DC 分布式无损智算推理网络技术验证，探索了 DeepSeek-671B 满血模型在 50km 拉远距离，不同输入序列长度、不同并发数、不同带宽收敛比下的分布式推理性能。本次试验基于 PD 分离架构，推理任务的 Prefill 阶段在边缘节点执行，部署少量算力（2 台服务器），模拟企业园区数据中心。推理任务的 Decode 阶段在云端执行，部署较多算力（4 台服务器），云、边之间通过 KV Cache 进行协同，有效降低推理服务成本，提升边缘节点的推理服务体验。试验中还对 DeepSeek 671B 满血版模型开展压力验证，系统评估了在 8:1、16:1、32:1、64:1 四种带宽收敛比下的推理拉远性能。

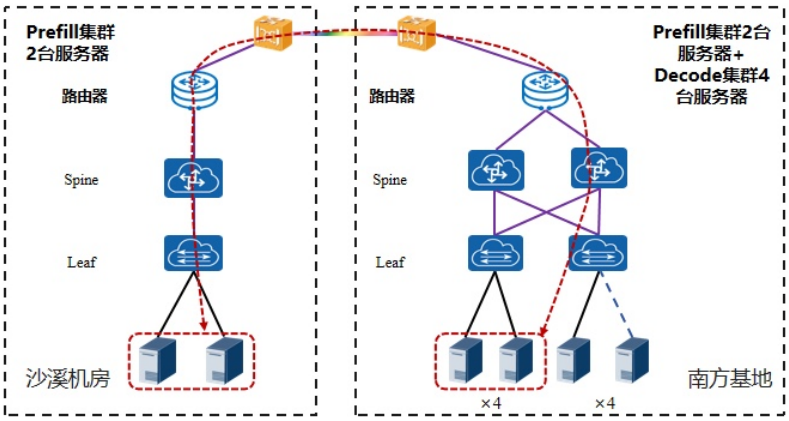


图 7-6 PD 分离推理拉远验证组网

3. 异构算力跨 DC 分布式推理

2025 年 7 月，项目组选取了业内不同 GPU 卡，进行异构算力 PD 分离推理拉远方案的测试验证。首先在本地组网中，采用 DeepSeek-R1 蒸馏模型，验证不同厂家 GPU 混合推理的技术可行性，通过不断改变推理参数，比如输入/输出序列长度组合、并发请求速率、最大并发数等，获取 TFTP、TPOT、Throughput 等推理性能指标。在此基础数据上，不断调整网络参数，进行拉远和非拉远推理

性能对比。拉远情况下网络调整的内容主要包括：OTN 拉远距离从 200km 延伸至 800km；长距链路的出口带宽逐渐收敛至 64:1，直至启动 ECN、PFC 流控机制；接入损伤仪，模拟网络的丢包、抖动，直至发生 RDMA 丢包重传。经过推理参数和网络参数的不断调整和组合，在 OTN 广域互联情况下，得到异构 GPU 间 PD 分离推理性能的变化规律，为今后现网部署提供了可靠的数据支撑。

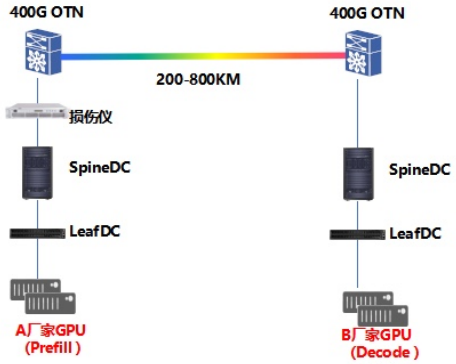


图 7-7 异构算力 PD 分离推理拉远验证组网

7.3.3 试验结论

项目组利用算间互联方案在全球首次解决了百公里长距跨 DC 大模型分布式训练难题。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，证明算间互联网络的可行性；网络稳定性方面，算间互联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。

项目组利用算间互联方案实现了企业算力资源的灵活扩容。试验结果表明，凭借算间无损网络低延迟、高吞吐的特性，结合 PD 分离架构设计，即使在带宽收敛比 64:1、百公里级长距拉远场景下，Deepseek 满血模型推理任务的 TTFT、TPOT、Throughput 三个关键指标均可以达到同 DC 场景的 99%，充分满足推理服务的 SLA 要求。而在异构算力推理拉远场景下，随着网络 QOS、拉远距离、收敛带宽等参数的调整，业务指标下降也可控制在 1%-5%以内，在保障推理性能的前提下，实现了异构存量算力的算力盘活。

8. 总结和展望

在新时代、新业态、新要求的背景下，智算业务呈现出算力需求激增、应用场景多元化等典型特征，对网络的性能、可靠性和适应性提出了更高要求。本白皮书围绕智算业务的典型需求和特征，深入研究了基于光电协同的智算网络相关方案、关键技术及典型实践。

在技术应用中，入算网络作为数据进入智算系统的第一道关口，借助光电协同提升了数据接入效率与安全性；算内网络通过光电协同架构重构了智算中心的网络范式，在端口级光交换、光电协同互联与集合通信优化中实现了多项技术突破，实现了大容量、超低功耗、 零时延、零拥塞、零人工干预的高可靠与弹性可扩展；算间网络则依托 IP 层技术与光传输层技术的协同优化，构建“算力-数据-网络”深度融合的互联体系，打破物理地域限制，实现算力、数据、存储资源的高效协同与池化调度，为跨城域甚至千公里级的分布式训练、推理奠定基础。这些技术的融合应用，构建起一套完整的光电协同智算网络体系，为 AI 大模型、超大规模智算服务和高可靠业务等场景提供坚实支撑，成为全球领先智算中心发展的网络内核。

未来，随着算力需求的持续增长，基于光电协同的智算网络技术将迎来更广阔的发展空间。其将进一步依托国家项目“面向分布式大模型的广域智算组网关键技术研究与应用示范”中的核心技术，针对OCS在不同网络架构中的部署位置、功能适配与性能优化开展更多创新性方案研究。同时，面向智算场景下预训练、微调、推理等多样性业务需求，构建“光网算用”一体化智算网络，并基于协同编排调度平台提供灵活、弹性的网络组网，实现智算资源池的可分可合。从而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
AIDC	AI Data Center	人工智能数据中心
APN	All Photonics Network	全光子网络
DBA	Dynamic Bandwidth Allocation	动态带宽分配
DPDK	Data Plane Development Kit	数据平面开发套件
DPI	Deep Packet Inspection	深度包检测
ECN	Explicit Congestion Notification	显式拥塞通知
FlexE	Flexible Ethernet	灵活以太网
FLOPS	Floating-point operations per second	每秒浮点运算次数
GPU	Graphics Processing Unit	图形处理器
ICMPv6	Control Message Protocol version 6	互联网控制消息协议第 6 版
I-ECN	Intelligent Explicit Congestion Notification	智能拥塞控制
LLM	Large Language Model	大语言模型
MEMS	Micro-Electro-Mechanical Systems	微机电系统
OCS	Optical Circuit Switch	光路交换机
ODU4	Optical Data Unit Level 4	光数据单元 4 级
OLT	Optical Line Terminal	光线路终端
ONU	Optical Network Unit	光网络单元
OSU	Optical Service Unit	光业务单元
OTU	Optical Transport Unit	光传送单元
OTUCn	Optical Transport Unit for Client	客户侧光传送单元
PD	Prefill-Decode	预填充解码
PFC	Priority-based Flow Control	基于优先级的流量控制
QP	Queue Pair	队列对
RDMA	Remote Direct Memory Access	远程直接数据存取
ROADM	Reconfigurable Optical Add-Drop Multiplexer	可重构光分插复用器
RoCE	RDMA over Converged Ethernet	基于融合以太网的远程直接内存访问
SDN	Software-Defined Networking	软件定义网络
SLA	Service Level Agreement	服务等级协议
TDM	Time Division Multiplexing	时分多路复用
TDMA	Time Division Multiple Access	时分多址
TTFT	Time To First Token	首令牌响应时延

附录 B：参考文献

[1] 中国电信股份有限公司研究院.分布式智算中心无损网络技术白皮书[R/OL](2024-8)

[2] Meta, A. I. "The llama 4 herd: The beginning of a new era of natively multimodal ai innovation." https://ai. meta. com/blog/llama-4-multimodal-intelligence/, checked on 4.7 (2025): 2025.

[3] Jouppi, Norm, et al. "Tpu v4: An optically reconfigurable supercomputer for machine learning with hardware support for embeddings." Proceedings of the 50th annual international symposium on computer architecture. 2023.

[4] Google Cloud Team. (2025). Ironwood (TPU v7): The First Google TPU for the Age of Inference. Google Cloud Blog


中国电信
CHINA TELECOM