智算中心光电混合组网下的流量差异化调度技术研究

Differentiated Traffic Scheduling for Hybrid Optical-Electrical Networking in Intelligent Computing Centers

韩博文/Han Bowen[1]，刘千仞/Liu Qianren[2]，
廖思忆/Liao Siyi[3]，胡雅坤/Hu Yakun[1]，曹畅/Cao Chang[1]
（1. 中国联通研究院，中国 北京 100048；
2. 中国联合网络通信集团有限公司，中国 北京 100033；
3. 中国联合网络通信集团有限公司上海市分公司，中国 上海 200082）
(1. China Unicom Research Institute, Beijing 100048, China;
2. China United Network Communications Group Corporation Limited, Beijing 100033, China;
3. China United Network Communications Group Co., Ltd. Shanghai Branch, Shanghai 200082, China)

DOI：10.12142/ZTETJ.202603005
网络出版地址：https://link.cnki.net/urlid/34.1228.TN.20260702.1533.001
网络出版日期：2026-06-26
收稿日期：2026-04-06

摘要：光电混合组网凭借“电交换的灵活可控”与“光交换的高带宽低时延”双重特性，已成为智算中心网络的核心发展趋势。聚焦智算中心光电混合组网场景，深入分析AI训练流量的差异化特征，并提出一种基于流量特征的差异化调度机制。该机制通过构建“感知-决策-执行-反馈”的闭环调度框架，实现了对数据并行（DP）流量、流水线并行（PP）流量等多种类型流量的精准识别与链路优势匹配。研究成果为智算中心光电混合组网下的流量差异化调度提供了理论依据与实践参考。

关键词：智算中心；光电混合组网；差异化调度；AI训练

Abstract: Optical-electrical hybrid networking, leveraging the dual advantages of flexible controllability of electrical switching and high-bandwidth low-latency of optical switching, is a key development trend in data center networks for artificial intelligence (AI). Focusing on the optical-electrical hybrid networking scenario in AI data centers, the differentiated characteristics of AI training traffic are analyzed in depth, and a traffic-characteristic-based differentiated scheduling mechanism is proposed. By constructing a closed-loop scheduling framework comprising perception, decision, execution, and feedback, accurate identification of different traffic types—including data parallelism (DP) and pipeline parallelism (PP) traffic—as well as matching with the advantages of optical or electrical links, is achieved. The findings provide both theoretical foundations and practical references for differentiated traffic scheduling in optical-electrical hybrid networks within AI data centers.

Keywords: intelligent computing center; optical-electronic hybrid networking; differentiated scheduling; AI training

引用格式：韩博文, 刘千仞, 廖思忆, 等. 智算中心光电混合组网下的流量差异化调度技术研究 [J]. 中兴通讯技术, 2026, 32(3): 22-29. DOI: 10.12142/ZTETJ.202603005

Citation: Han B W, Liu Q R, Liao S Y, et al. Differentiated traffic scheduling for hybrid optical-electrical networking in intelligent computing centers [J]. ZTE technology journal, 2026, 32(3): 22-29. DOI: 10.12142/ZTETJ.202603005

数字经济与人工智能（AI）的深度融合，正推动全球AI算力基础设施建设进入高速扩张周期。AI大模型的训练与推理应用导致算力需求呈现几何级增长。以Seedance 2.0为代表的多模态大模型及OpenClaw等开源AI智能体的广泛部署，进一步加速了算力需求的攀升。在此背景下，算力供给能力已成为衡量国家数字经济核心竞争力的关键指标。2026年，中国政府工作报告[1]首次将“超大规模智算集群”列为新基建工程。这一举措不仅从国家战略层面确立了智算基础设施的核心地位，同时也对其底层网络的支撑能力、资源调度效率与整体能效提出了前所未有的严苛要求。

随着智算集群规模持续扩大，超大规模AI数据中心的发展瓶颈已从单芯片算力逐步转向集群级互连带宽、能耗密度与系统可扩展性。在此条件下，网络基础设施成为制约智算集群算力充分释放的核心因素。在大规模智算集群的组网架构中，传统纯电交换网络虽具备调度灵活的优势，但在超大带宽扩展、超低时延传输和能耗控制方面存在明显短板；

纯光交换网络虽拥有高带宽、低时延和低功耗的特性，却面临突发流量适配能力差、与现有电交换系统兼容性不足的问题，从而限制了其规模化应用。

作为智算网络技术演进的重要风向标，2026年全球光通信大会（OFC）与图形处理器（GPU）技术大会（GTC）为突破上述瓶颈提供了明确的技术路径。在OFC大会上，Google、Nvidia等厂商展示了光电路交换机（OCS）在张量处理器（TPU）/GPU集群中的规模化应用案例，证实光电混合架构是突破传统电交换网络带宽与功耗瓶颈的重要方向；在GTC 2026上，Nvidia发布了Spectrum-X共封装光学（CPO）交换机，并提出“光铜并行、光电混合”的技术战略。

在此背景下，融合“电交换灵活可控”与“光交换高带宽低时延”双重优势的光电混合组网方式，成为突破智算中心网络瓶颈、支撑下一代高性能智算集群的核心发展趋势。其中，差异化流量调度技术作为光电混合组网的关键支撑，通过为不同智算业务设计定制化调度策略，实现光/电网络资源的协同优化与动态匹配，是提升组网整体性能、释放超大规模智算集群算力的重要途径。基于上述分析，本文聚焦智算中心光电混合组网场景，开展差异化调度技术研究，以期突破现有调度模式的局限，为提升智算中心网络资源利用率与算力释放效率提供技术参考。

1 全球研究现状

1.1 Google

Google作为智算集群光电混合组网技术的先行者，其Jupiter组网架构与第7代自研TPU芯片（TPUv7，代号Ironwood）的深度协同，构建了面向超大规模AI大模型训练的高效网络解决方案，为光电混合组网及调度技术提供了重要实践参考[2]。

Jupiter架构初期采用三级Clos拓扑，由机架顶交换机（ToR）、聚合块及脊层块层级互联构成，然而该拓扑难以适配超大规模数据中心的增量扩展需求与智算负载的通信特征，主要瓶颈体现在4个方面：

1）Spine层预部署导致带宽降级：Spine层需一次性预部署，后续新代际聚合块的高速链路会因Spine层低速链路被降级，造成算力与网络带宽失衡；

2）增量升级成本高、易中断：Spine层升级需全网重新布线，操作复杂且影响生产流量，无法实现计算/存储硬件的渐进式更新；

3）功耗与成本冗余：Spine层交换机及配套光模块带来额外的资本支出（CAPEX）和功耗，且链路利用率低；

4）无法适配异构速率：不同代际如100G、200G或400G链路难以共存，无法匹配AI集群的差异化带宽需求。

2022年，Google提出Jupiter架构的核心演进方案[3]（如图1）。该方案通过引入基于微机电系统（MEMS）的OCS，构建了数据中心互联层（DCNI），能够为数据中心网络创建任意逻辑拓扑，实现动态拓扑重构；同时，通过集中式软件定义网络（SDN）控制，进行流量工程优化与自动化网络运维。该演进方案推动Jupiter架构实现了从传统Clos拓扑向块级直连拓扑的范式升级。相较原Clos拓扑，网络带宽与容量提升5倍，功耗降低41%，CAPEX降低30%，充分验证了光电混合组网技术在超大规模数据中心及智算集群中应用的可行性与工业价值。

Google自研的TPU是面向AI大模型训练与推理的专用加速芯片。历经多代技术迭代，TPU已形成规模化集群部署能力。在Jupiter光电混合组网架构的基础上，Google结合TPU自身的芯片互联特性进行定制化创新，打造了适配超大规模智算需求的TPU专属集群组网架构与调度策略。

TPU集群组网的核心思路为“光电分域、分层互联、弹性扩展”，其整体架构如图2所示。该架构以4×4×4 Cube为最小算力组网单元，构建“单元内电交换、单元间光交

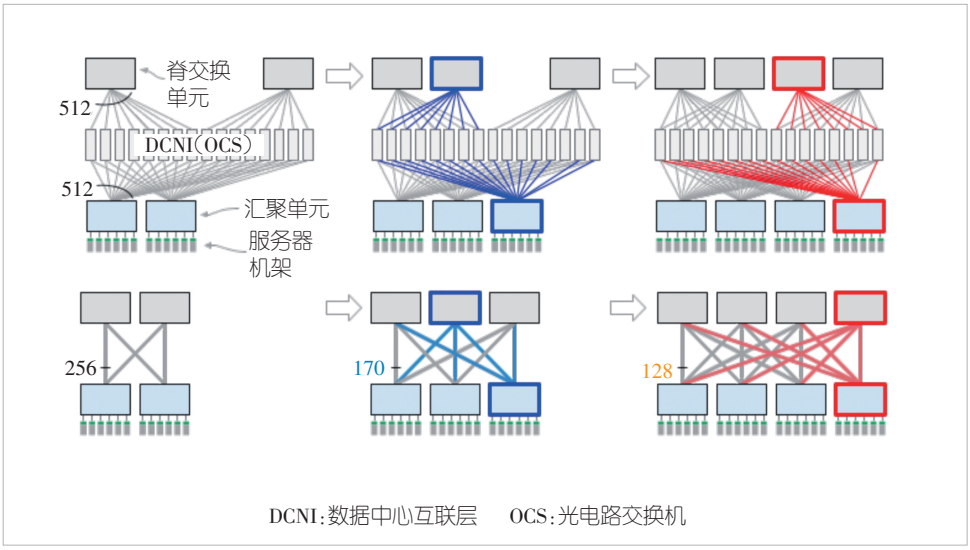


图1　基于DCNI层支持Jupiter架构渐进式扩展

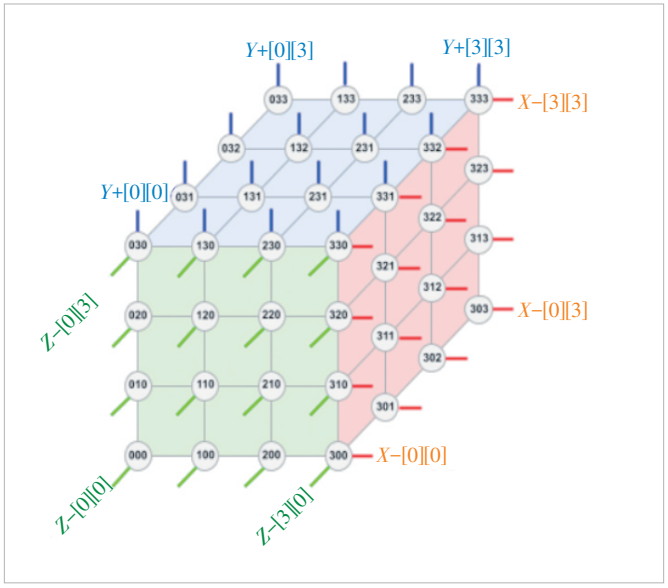


图2　TPU Cube示意图

换”的分层结构：Cube内部通过芯片间互联链路实现全电信号紧耦合通信；多个Cube之间则基于OCS构建三维Torus（3D Torus）拓扑。OCS按三维正交方向部署，支持毫秒级的动态拓扑重构。

TPU集群组网调度策略深度贴合AI大模型训练的通信特征，通过算力切片实现集群资源的弹性分配与多租户隔离；针对稠密/稀疏计算的差异化流量实施分层调度，匹配拓扑高带宽维度与优化拓扑；围绕数据、模型、流水线并行策略动态重构光拓扑，适配不同并行方式的通信需求；依托OCS的快速光路切换能力实现故障自愈，在不中断训练任务的前提下完成流量重路由，最终实现算力与网络资源的精准匹配，最大化释放TPU集群的算力性能。

该调度策略深度贴合AI大模型训练的通信特征，具体体现在5个方面：

1）通过算力切片实现集群资源的弹性分配与多租户隔离；

2）针对稠密计算与稀疏计算的差异化流量实施分层调度，匹配拓扑的高带宽维度，并优化网络拓扑结构；

3）围绕数据并行、模型并行与流水线并行策略，动态重构光拓扑，以适配不同并行方式的通信需求；

4）依托OCS的快速光路切换能力实现故障自愈，在不中断训练任务的前提下完成流量重路由；

5）最终实现算力与网络资源的精准匹配，最大化释放TPU集群的算力性能。

1.2 Nvidia

在2023年OFC上，Nvidia针对传统数据中心网络物理布线固化、故障自愈机制存在明显缺陷的问题开展了研究。现有故障恢复方案多依赖路由重调整或冗余硬件部署，不仅易导致集群性能降级、硬件资源利用率低下，而且无法解决叶交换机故障引发的服务器断连问题。

针对上述不足，Nvidia提出将SDN控制能力下沉至物理层（L1）的技术方案[4]，如图3所示。该方案基于OCS构建

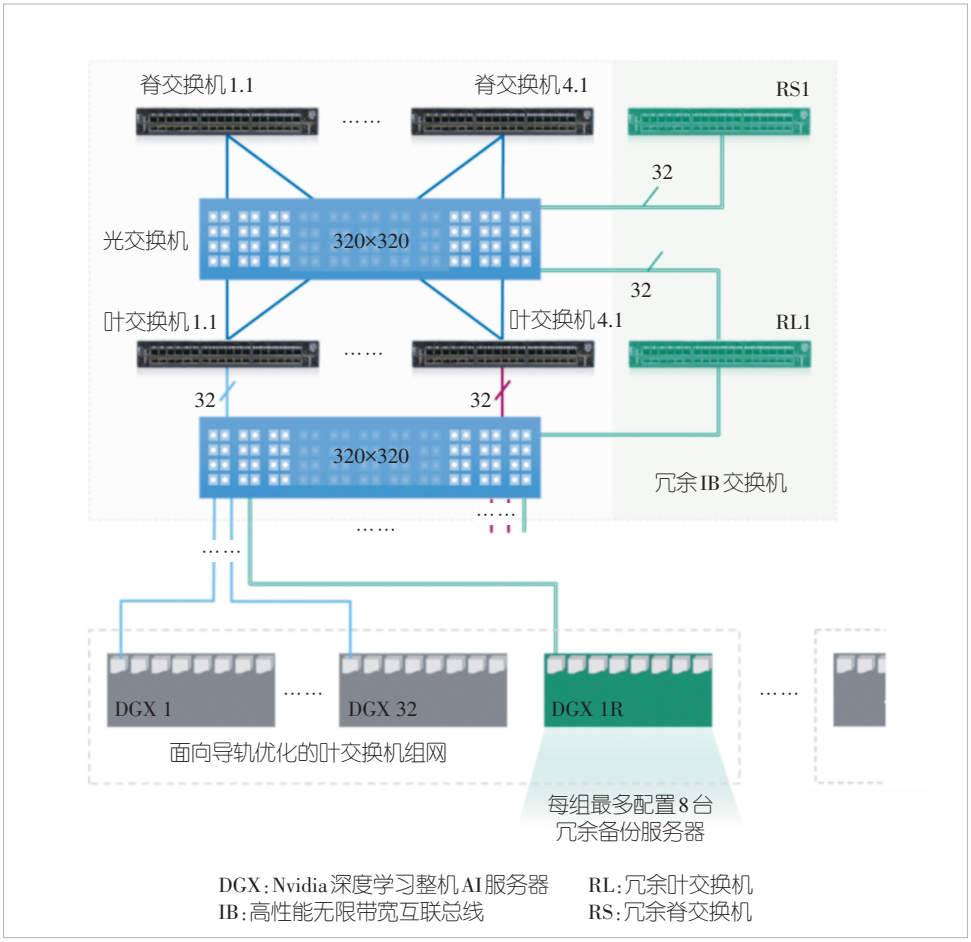


图3　高可用架构图

可编程光交换矩阵，在两层胖树（Fat-tree）拓扑中嵌入OCS并部署冗余叶/脊交换机。同时，通过自研光交换管理器（OFM）实现对OCS的集中控制，完成故障设备的秒级物理链路切换与拓扑重构。

该方案在由4台DGX服务器与14台InfiniBand（IB）交换机构建的测试床中进行了实验验证。结果表明，当叶/脊交换机发生故障时，集群可在数秒之内恢复至满性能状态，有效避免了因叶交换机故障导致的应用崩溃。该方案从根本上解决了传统路由自愈方案中存在的性能降级与冗余硬件利用率低的问题。

总体而言，该方案实现了物理层拓扑的动态可编程，以轻量化冗余硬件达成了高效的故障自愈，为高性能计算（HPC）及AI集群的网络高可用设计提供了全新的技术路径。

2024年，Nvidia在2023年技术成果的基础上，针对大模型预训练集群中存在的三大痛点——传统多级胖树拓扑固化、功耗与成本高昂、难以适配任务通信特征，提出了基于OCS的拓扑动态适配方案[5]。

该方案利用OCS层替代传统Fat-tree架构中的高层电交换结构。具体实施路径如下：首先，依据大语言模型（LLM）、深度学习推荐模型（DLRM）等典型任务的预期通信模式，通过资源管理工具完成算力与网络资源的分配；随后，利用OCS动态构建环形、全连接等适配物理拓扑，实现网络架构与业务负载的深度适配。

在性能表现方面，该方案不影响单任务的带宽资源，集群资源利用率与传统Fat-tree架构的差距控制在1%以内（如图4）。同时，该方案可降低网络功耗超过50%、网络成本超过30%，为超大规模AI训练集群提供了高效低耗的组网支撑。

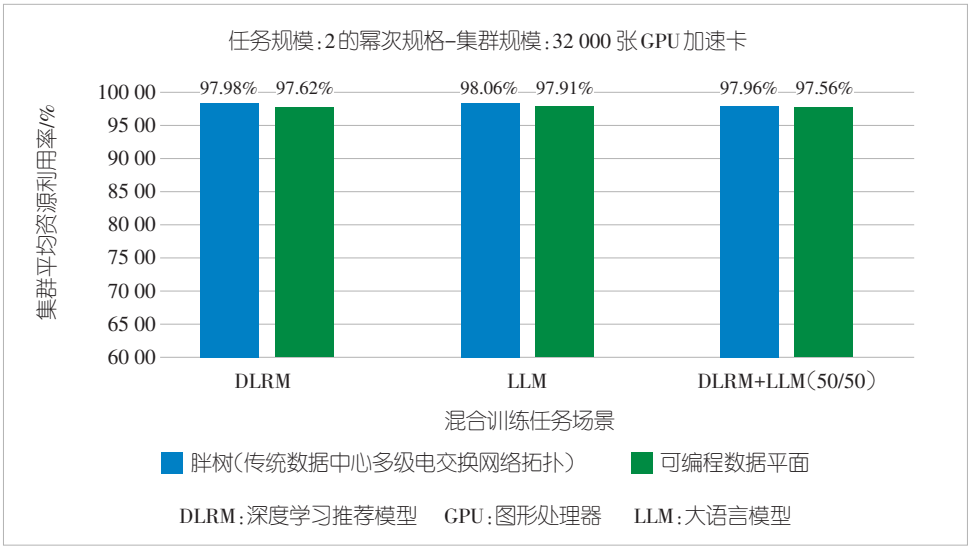


图4　基于DCNI层支持Jupiter架构渐进式扩展

1.3 百度

针对Meta提出的“Rail-only”单导轨组网架构，百度提出了以OCS替换二层以太网分组交换平面（EPS）的改进方案[6]，如图5所示。该方案将每个TOR组内的流水线并行（PP）流量控制于单台TOR交换机所辖的GPU之间，并将每个Group中编号相同的TOR交换机上联至同一个OCS交换平面。

同时，为支持全互连（all-to-all）通信，百度进一步提出了基于OCS对大规模二层EPS平面进行扩展的网络架构，如图6所示。在该架构中，每个机架顶（ToR）交换机需拆分出部分端口用于连接相应的OCS平面，同时将剩余端口连接至叶交换机。该设计既保留了每个二层EPS平面内的全互连通信能力，又具备了规模扩展以支持更大规模训练任务的能力。

2 流量差异化调度机制

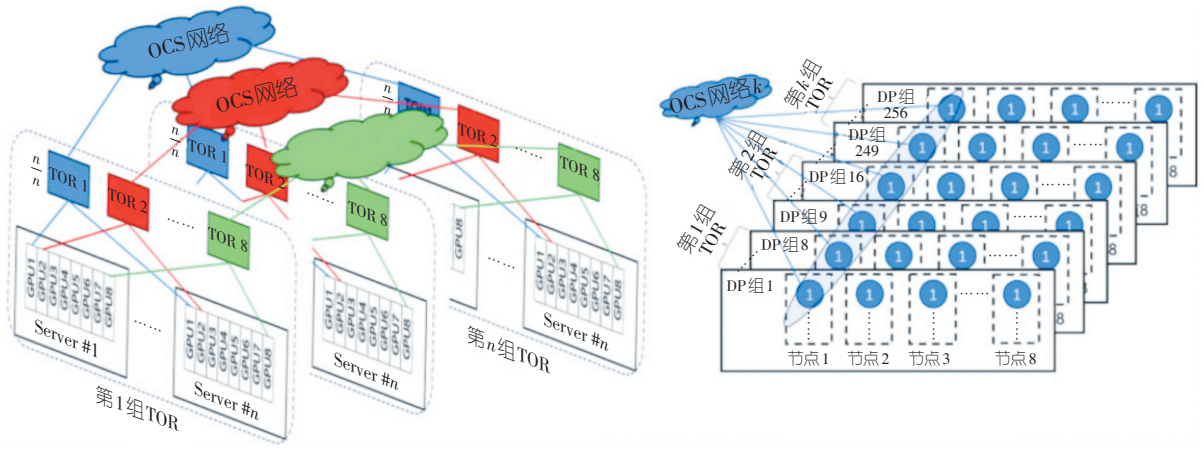
2.1 面临挑战

在光电混合组网场景下，流量调度面临以下核心挑战：

1）流量特征识别与适配难度大。AI训练等高性能计算场景中的流量具有“突发式、高并发、多类型”的特征。例如，数据并行（DP）流量为大带宽、高并发的全局同步流量，流水线并行（PP）流量为低时延、顺序敏感的点对点流量。如何精准识别不同类型流量的特征，并针对性地匹配光链路与电链路的各自优势，是调度面临的核心难点。

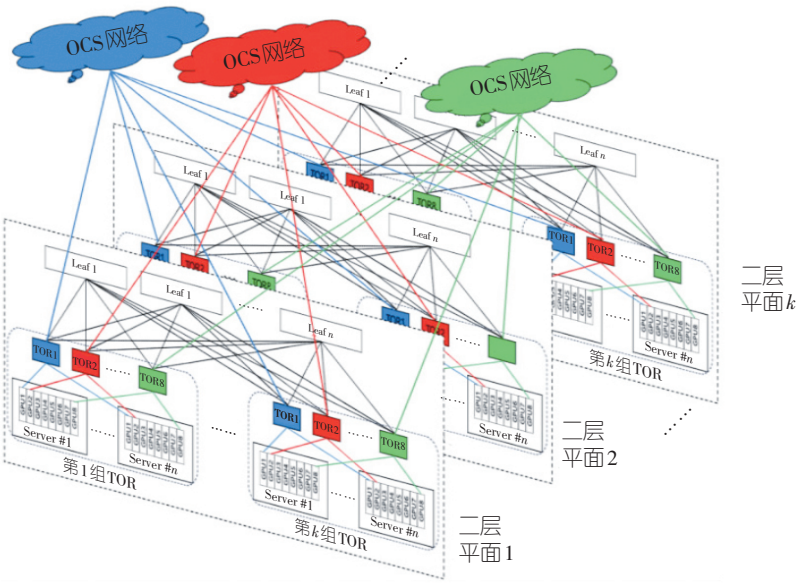
2）链路协同与负载均衡难题。光链路与电链路在带宽、时延、转发机制等方面存在固有差异，传统等价多路径（ECMP）负载均衡算法难以直接适配，容易导致链路负载不均。光链路带宽虽高但灵活性不足，若流量调度不当，可能出现光链路闲置而电链路拥塞的状况。同时，光链路的切换延迟可能引发流量中断，进而影响业务连续性。

3）调度策略的动态适应性不足。现有调度方案多基于预设规则或简单的实时流量监测，难以



DP：数据并行　OCS：光电路交换机　TOR：机架顶端接入交换机

图5　基于DCNI层支持Jupiter架构渐进式扩展



OCS：光电路交换机　TOR：机架顶端接入交换机

图6　大规模二层EPS平面基于OCS进行扩展

应对复杂场景下流量的动态变化，从而导致网络性能下降。此外，不同业务场景（如训练、推理）的流量特征差异较大，通用调度策略难以兼顾各类场景的差异化需求。

4）拓扑约束与扩展性限制。基于OCS的光交换虽可实现毫秒级或秒级的物理层光路切换与拓扑动态调整，但在大规模集群中，若所有节点间均通过OCS实现全互连，所需端口数与连接数将随节点规模呈平方级增长，导致光路调度与维护复杂度显著上升。

5）运维与验证复杂度高。光电混合组网涉及光交换机、电交换机、SDN控制器、业务平台等多设备、多系统的协同，调度策略的配置、部署与验证流程较为复杂。具体而言，如何验证调度策略的有效性、如何定位光/电链路切换过程中的故障、如何评估调度方案对业务性能的实际提升效果，均需依赖专业的运维工具和测试方法，这进一步增加了运维成本。

2.2 设计框架

本方案的核心设计思路为“流量特征驱动、链路优势匹配、动态智能调度”，即针对AI训练场景中的DP、PP等核心流量类型，构建差异化调度机制，以实现光链路与电链路资源的最优利用。

方案采用“感知–决策–执行–反馈”的闭环设计框架。该框架包含以下4个核心模块：

1）感知模块。该模块包括模型切分策略感知与网络状态感知两个子功能。

（1）模型切分策略感知：获取AI训练任务的参数配置，包括张量并行（TP）/PP/DP切分方式、秩（rank）划分、全局批大小（GBS）设置等，为调度决策提供数据支撑。

（2）网络状态感知：通过管控平台获取智算网络的链路层发现协议（LLDP）拓扑连接关系（包括交换机与智算服务器之间、交换机与交换机之间的拓扑连接），同时利用Telemetry遥测功能实现对网络流量的毫秒级监测。

2）调度决策模块。该模块基于感知模块采集的数据，构建差异化调度规则库。核心规则如下：

（1）DP流量调度：DP流量是AI训练中带宽需求最大、对性能影响最为关键的流量类型，其调度机制是本方案的核心内容。

a）流量识别：通过网络拓扑连接关系与AI训练参数（DP切分方式、rank划分等），可预先计算出DP流量的通信关系，进而实现流量识别。

b）链路选择优先级：第1优先级为OCS光链路，第2优先级为光电混合链路（采用ECMP负载分担），第3优先级为纯电链路。

（2）PP流量调度：PP流量具有低时延、顺序敏感、数据量相对较小的特点。调度策略根据链路负载动态选择光链路或电链路——若光链路负载较低，则优先选择光链路；若光链路负载过高，则切换至电链路。

3）执行模块。该模块通过管控平台将调度决策转换为具体的配置指令，下发至光交换机与电交换机。例如，通过配置访问控制列表（ACL）规则实现流量的链路定向转发，或通过边界网关协议（BGP）调整路由策略以实现链路切换。

4）反馈优化模块。该模块实时监测调度执行后的网络性能（如链路负载、时延、带宽利用率）与业务性能（如AI训练的每秒浮点运算次数（TFLOPS）、迭代耗时），并将监测结果反馈至调度决策模块。若网络性能或业务性能未达预期，则动态调整调度规则，优化链路分配策略。

3 试验验证与结果分析

3.1 试验环境与测试方案

试验环境采用典型的Spine-Leaf两层架构，在Spine层引入异构OCS设备，逻辑拓扑如图7测试拓扑所示。

1）计算层：部署4台AI服务器，每台服务器搭载8张AI加速卡（共32张GPU），单台服务器配置8个200G远程直接内存访问over融合以太网（RoCE）网卡，单机总带宽为1.6 Tbit/s，构成一个标准的训练集群。

2）接入层：每台Leaf交换机下行连接一台服务器上的8个网卡，上行通过400G链路连接至Spine层。上行链路中，一半连接至电交换机，另一半连接至OCS。Leaf交换机的上下行带宽保持1∶1无收敛配置。

3）核心层：部署1台电交换机和1台OCS光交换机。每台逻辑Leaf交换机同时上行连接至电Spine与光Spine，通过BGP或静态路由控制流量的转发路径。

3.2 光电性能对比测试

在双机（16卡）训练场景下，针对二层只走电交换机或二层只走光交换机的场景，我们测试了Llama2（6.7B）模型的训练性能，逻辑拓扑如图8所示。

由于节点数少，网络拓扑结构相对简单，流量冲突概率低。测试结果如表1所示。

测试结果表明：在小规模、无阻塞的简单拓扑中，OCS完全可以替代电交换机，且能提供更优的时延表现。

3.3 调度策略对比测试

在32卡集群环境下，我们开展了Llama2（34B）模型的训练性能对比实验，分别评估了不同调度策略下的性能表现。同时，通过调整全局批大小（GBS），改变了训练过程中的“计算通信比”，从而放大不同网络架构之间的性能差异。表2展示了32卡环境下、不同GBS取值时的训练性能对比结果（TP=8，PP=2，DP=2）。

我们对测试结果进行了分析：

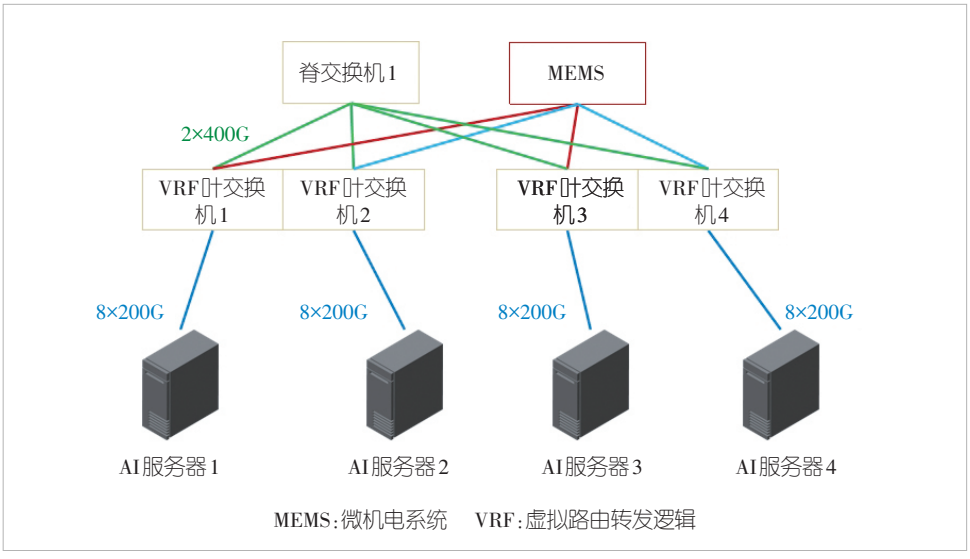
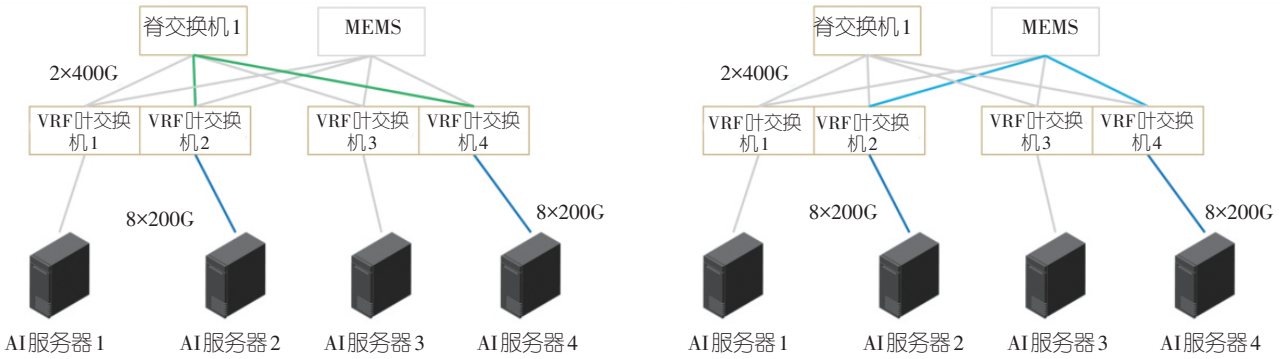


图7　测试拓扑



MEMS：微机电系统　VRF：虚拟路由转发逻辑

图8　16卡训练示意图

表1　双机训练场景性能对比

配置场景	网络路径	平均耗时/ms	平均TFLOPS
TP8 PP1 DP2	电交换（spine1）	72 076.36	73.53
TP8 PP1 DP2	光交换（mems）	72 061.29	73.56
TP8 PP2 DP1	电交换（spine1）	78 087.56	67.87
TP8 PP2 DP1	电交换（mems）	77 340.46	68.52

DP：数据并行　TFLOPS：每秒万亿次浮点运算
PP：流水线并行　TP：张量并行

表2　不同GBS场景下的性能对比

GBS	DP走光（TFLOPS）	PP走光（TFLOPS）	性能差异/%
2	25.82	22.53	14.6
16	86.14	78.94	9.1
32	100.19	96.66	3.7
64	111.39	108.92	2.3
128	117.26	114.87	2.0

DP：数据并行　PP：流水线并行
GBS：全局批大小　TFLOP：每秒万亿次浮点运算

1）当GBS=2时，系统实际算力仅能达到约25 TFLOPS，约为峰值算力的25%。此时网络通信极为频繁，协议栈处理开销被明显放大。即便如此，DP流量经光链路传输的方案（25.82 TFLOPS）仍优于PP流量经光链路传输的方案（22.53 TFLOPS），性能提升幅度达14.6%。该结果验证了在高频通信压力下，将DP流量优先调度至低时延OCS链路的有效性。

2）GBS=16处于典型的性能“过渡区”，此时计算与通信的时间开销大致相当，网络调度策略的选择对最终性能影响最大。DP流量经光链路传输的方案较PP流量经光链路传输的方案性能高出9.1%。该结果进一步验证了在光电混合架构中，将DP流量优先调度至OCS链路的核心策略。

3）当GBS增大至128时，DP流量经光链路与PP流量经光链路的性能差距缩小至2%。此时GPU绝大部分时间处于满负荷计算状态，网络通信时间占比极低，网络架构选择的容错率显著提高。

4 总结与展望

本文聚焦智算中心光电混合组网场景，针对AI训练中流量的差异化调度需求，提出了一种基于流量特征的差异化调度机制。该机制通过构建“感知—决策—执行—反馈”的闭环调度框架，实现了对DP流量、PP流量等不同类型流量的精准识别，并据此完成光链路与电链路的优势匹配。

未来工作将针对混合专家模型（MoE）及推理场景进行专项优化，以进一步提升流量识别精度与调度策略的动态适应性，从而推动光电混合组网技术在智算中心中的持续发展与应用。

参考文献

[1] 中华人民共和国国务院. 2026年政府工作报告 [R/OL]. (2026-03-13) [2026-04-07]. http://www.gov.cn/yaowen/liebiao/202603/content_7062620.htm
[2] 叶通, 胡卫生. 大规模智算中心光电交换网络架构演化综述 [J]. 电信科学, 2025, 31(4): 32-44. DOI: 10.11959/j.issn.1000-0801.2025116
[3] Poutievski L, Mashayekhi O, Ong J, et al. Jupiter evolving: transforming Google's datacenter network via optical circuit switches and software-defined networking [EB/OL]. [2026-04-10]. https://web.stanford.edu/class/cs244/papers/poutievski-sigcomm22.pdf
[4] Patronas G, Syrivelis D, Bakopoulos P, et al. Software-defined, programmable L1 dataplane: demonstration of fabric hardware resilience using optical switches [C]//2023 Optical Fiber Communication Conference (OFC). IEEE, 2023: Th2A.15
[5] Bakopoulos P, Patronas G, Terzenidis N, et al. Photonic switched

networking for data centers and advanced computing systems [C]//2024 Optical Fiber Communication Conference (OFC). Optica Publishing Group, 2024: M2G.1
[6] 朱宸, 周谐, 王佩龙. 可重构OCS技术在大模型预训练中的应用 [J]. 光通信研究, 2024 (5): 29–38. DOI:10.13756/j.gtxyj.2024.240049

作 者 简 介

韩博文，中国联通研究院高级研究员，工程师；主要研究领域为智算网络关键技术、管控调度技术等。

刘千仞，中国联合网络通信集团有限公司建设发展部高级工程师；主要研究领域为云计算、算力网络等。

廖思忆，中国联合网络通信集团有限公司上海市分公司云网运营中心高级总监；主要研究领域为下一代互联网、算力网络、数据中心网络等。

胡雅坤，中国联通研究院光传输领域研究员；从事超高速传输与智算光互联方向的理论研究与工程化应用，深度参与中国联通算力智联网建设工作，围绕智算内和智算间光互联开展技术创新。

曹畅，中国联通研究院下一代互联网研究部总监，正高级工程师；主要研究领域为算力网络、IPv6+网络新技术、未来网络体系架构等。