智算中心光电协同交换网络全栈技术白皮书

湖南大学 中国联通研究院
中国联通软件研究院 北京邮电大学 上海交通大学

2025 年 8 月

编写说明

编写单位：

湖南大学、中国联通研究院、中国联通软件研究院

北京邮电大学、上海交通大学

编写人员：

湖南大学：

陈果、梁帮博、陈禹澎、刘璇

中国联通研究院：

程新洲、曹畅、徐博华、杨斌、文晨阳、谢志普、徐洁、黄金超

中国联通软件研究院：

杨迪、李张体、张承琪、王宇、马煜

北京邮电大学：

邢颖、林雪燕

上海交通大学：

赵世振

前言

人工智能正以前所未有的速度重塑人类生产与生活方式。以大语言模型、多模态模型为代表的新一代 AI 应用，持续突破计算与通信的极限，推动智算中心从计算、存储到网络的全栈架构深度演进。在这一浪潮中，智算中心不仅是国家科技战略的核心支撑，更是产业智能化升级的关键基础设施。

随着 AI 模型参数量呈指数级增长，尤其是在大规模分布式并行训练场景下，网络性能已成为制约智算中心整体效率的关键瓶颈。当前普遍部署的纯电交换网络在互联规模、带宽密度、端到端时延与能效比等方面逐渐逼近物理与经济的上限：算力芯片的通信需求远超传统网络承载能力，高功耗、高成本和复杂布线问题愈发突出。

在此背景下，光交换技术凭借超大带宽、超低延迟与低功耗等特性，正与电交换形成互补融合的“光电协同”架构，成为新一代智算中心网络的重要发展方向。光电协同不仅能够在物理层显著提升链路性能，还为网络的灵活重构、智能调度与按需适配提供了技术空间。全球领先的产业与科研力量均已在此领域展开探索，并在部分应用场景实现试点部署。

然而，要实现光电协同网络在智算中心的规模化落地，仍需跨越多重技术关卡。从应用层集合通信模式与动态拓扑的适配，到传输层协议机制与流量调度优化；从路由层控制平面的可扩展性，到链路层资源的智能分配；再到物理层光交换的传输损耗与延迟难题，均对网

络架构设计、协议栈演进与资源编排提出了系统性挑战。

本白皮书面向智算中心光电协同交换网络的全栈技术体系，旨在：

- 梳理国家政策、AI 发展趋势与智算中心网络需求，揭示光电协同兴起的背景；
- 分析光交换与电交换的性能差异与技术互补性；
- 总结光电协同网络在应用层、传输层、路由层、链路层与物理层的关键挑战与发展路径；
- 提出面向未来的技术演进方向与标准化路线建议。

我们期望本白皮书能为智算中心网络领域的研究人员、设备制造商、运营商与服务提供商，提供系统的参考框架与技术洞察，共同推动构建超大规模、超大带宽、超低时延、超高可靠的新一代智算中心网络基础设施。

本白皮书的编制工作得到了国家自然科学基金项目（编号：U24B20150）的支持，在此表示感谢。

目录

1. 智算中心发展与光电协同交换网络兴起

1.1 国家政策发展

全球智能化浪潮风起云涌，人工智能领域创新呈突破之势，语言大模型、多模态大模型和具身智能等领域日新月异，推动以智算中心为代表的基础设施向更高效、更弹性的方向快速发展。

2025 年 1 月 1 日，国家发展改革委等联合印发《国家数据基础设施建设指引》[1]强调高效弹性传输网络可为大模型训练和推理等核心场景数据传输流动提供高速稳定服务，在高效弹性传输网络支撑下，能够显著提升数据交换性能，降低数据传输成本。

7 月 26 日，李强总理出席 2025 世界人工智能大会暨人工智能全球治理高级别会议开幕式，围绕如何把握人工智能公共产品属性、推进人工智能发展和治理发表致辞。大会发表《人工智能全球治理行动计划》[2]协力推进全球人工智能发展与治理。该计划指出应“加快数字基础设施建设”，即加快全球清洁电力、新一代网络、智能算力、数据中心等基础设施建设，完善具备互操作性的人工智能和数字基础设施布局，推动统一算力标准体系建设。

这些政策举措充分体现了我国在人工智能基础设施建设方面的前瞻性布局，通过政策引导、标准制定和国际合作等方式，为人工智能技术创新和产业发展构建坚实的算力支撑体系，同时为智算中心的快速发展注入了强大的助推剂。

1.2 智算中心发展

据中国互联网络信息中心的报告[3]，2024 年我国人工智能产业规模突破 7000 亿元，连续多年保持 20%以上的增长率。2025 年上半年，生成式人工智能产品实现了从技术到应用的全方位进步，产品数量迅猛增长，应用场景持续扩大。

在人工智能+医疗领域，医联 MedGPT、神农中医药大模型和岐黄问道等医疗大模型已广泛应用于辅助诊断、中医诊疗、智能开方等环节，显著提升了医疗服务质量和效率。在人工智能+汽车领域，大模型推动变革汽车产业全链条，全面智能化升级。华为盘古汽车大模型聚焦汽车产业全链条场景，覆盖设计、生产、营销、研发等核心环节，为汽车行业垂直领域解决方案。在数据要素价值释放过程中，强大的算力可以将“大数据”转向“好数据”，并充分挖掘海量数据的经济和社会价值，不断激活数据要素潜能，实现原始数据向知识再向智慧跃迁的更高层次价值释放。

随着人工智能与实体经济深度融合，智算需求已经呈现爆发式增长。AIGC 大模型参数量达到万亿，训练阶段需要万卡甚至十万卡集群支持。如表 1-1 所示，训练万亿级模型（如 GPT-4）已突破万亿（10^{25}）FLOPs，需数千至万块 H100 级芯片，训练成本达上亿美元。

表 1-1 不同规模模型的算力需求估算

模型规模	典型硬件	GPU 数量	训练成本（美元）
1 亿~10 亿	V100/A100（数十卡）	<100	<10k
百亿级	A100/H100（数千卡）	1,000-5,000	1M-10M
千亿级	H100（万卡级）	5,000-10,000	10M-100M
万亿级+	H100/B100（数万卡）	>20,000	100M-500M+

大模型参数量达到万亿，迭代训练需使用数据并行、流水线并行、张量并行和专家并行等技术。并行推理将每个模型层的计算任务拆分到各个服务器中多卡 GPU 上执行。各 GPU 无法独立完成计算工作。在训练的过程中需要进行频繁且复杂的通信。这就要求构建 GPU 之间的全互联高速数据通道，以确保数据的高效传输，最大限度减少 GPU 间通信耗时。那么，如何满足大规模 GPU 之间的高效通信，构建超大规模、超大带宽、超低时延、超高可靠的智算网络，已成为当前智算网络发展重要挑战。

智算中心网络如图 1-1 所示，可按通信范围分为**机内互联（Intra-Node）**与**机外互联（Inter-Node）**两类：

机内互联：主要用于单服务器或单节点内的多 GPU 连接。典型技术包括 PCIe 与 NVLink，其中最新一代 NVLink[4] 5.0 点对点带宽高达 1800 GB/s，并通过 NVLink Switch 实现多 GPU 全互联，支持构建大规模 GPU 池。

机外互联：用于跨服务器或跨机柜的 GPU 通信，需依赖高速网络结构实现。当前主流方案采用电交换芯片构建以太网或 IB 网络，常见架构包括 Fat-Tree、Leaf-Spine、DCell、BCube。这些结构通过

多层交换机实现大规模互联，支撑分布式训练中的全互联需求。

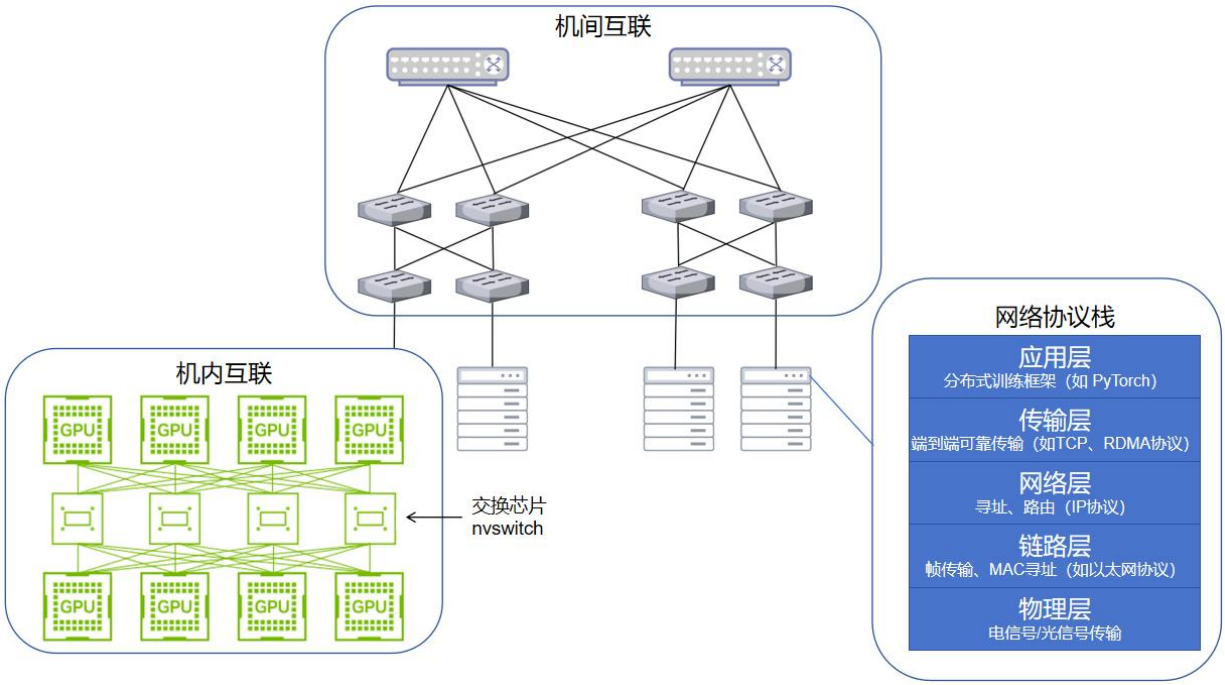

图 1-1 智算中心网络与网络协议栈

无论采用机内互联还是采用机外互联，都要采用电交换芯片来做网络流量交换。然而，随着模型规模和节点数的增加，电交换面临带宽、延迟和能效的瓶颈。

1.3 光电协同交换网络的兴起

在交换技术方面，电交换技术具有成熟性、协议兼容性和灵活的控制能力，基于以太网（如 RoCEv2、InfiniBand）传输协议，支持复杂网络策略，在智算中心广泛部署。基于电交换机的典型的架构包括 Fat-Tree、Leaf-Spine、Dcell、BCube 等。受限于集成电路工艺的发展限制，传统电交换机的带宽密度已难以满足大模型训练增长的流量需求。光交换具有大带宽、可靠性高、功耗小、组网灵活的特点，相比电交换机具有高带宽、低能耗的优势，是突破网络核心侧带宽密度瓶颈的最佳技术路线，适用于超大规模 AI 训练集群。光电协同架构[6]可以将光交换的高带宽、低延迟和电交换的灵活控制能力整合起来，提供 TB 级带宽，充分发挥光与电两者优势。

表 1-2 光电交换技术比较

	光电协同	全电交换	全光交换
带宽	TB 级	≤800Gbps	TB 级
延迟	纳秒（光）+微秒（电控制）	微秒级	纳秒级
能效	功耗较低	功耗高	功耗低
成本	一般	较低	较高

1.3.1 电交换的技术瓶颈与发展困境

端口密度瓶颈

尽管近年来电交换芯片在制程工艺、转发架构与缓存设计方面不断优化，但在智算中心应用场景下，其性能仍面临明显瓶颈。随着摩尔定律逐渐失效，交换芯片的更新迭代速度明显放缓，芯片交换容量难以实现持续增长。目前主流商用电交换芯片已发展至 102.4 Tbps 级别，例如 Broadcom Tomahawk 6 采用 3nm 制程工艺，可提供多达 128 个 800 G 端口或 64 个 1.6T 端口。而国产交换芯片仍停留在 7nm 制程的 25.6Tbps 交换容量，瓶颈效应更加严重。然而在实际部署中，为保障链路冗余、流控带宽和管理接口，芯片可用端口通常不到理论最大值，导致整体带宽扩展能力受到压制。尤其是在并行训练中伴随的突发性大量同步与广播时，网络时常出现瞬间拥塞、缓存溢出与延迟剧增等问题[7]。

与此同时，随着大模型参数规模和训练复杂度的持续增长，智算中心对网络端口密度的需求正加速攀升。以 GPT-4 等万亿级模型为例，其完整训练任务需部署约 25,000 张 H100 GPU 卡。假设每台服务器需与 Top-of-Rack（ToR）交换机建立至少 2 条 400G 上行链路，并在 Leaf 层与 Spine 层交换节点之间形成全互联结构，则光是 Leaf 层汇聚这些服务器所需的交换芯片就需提供数千个高带宽端口。进一步向上扩展 Leaf 层与 Spine 层的连接关系时，每增加一层交换所需的端口数将指数增长，极易突破现有交换芯片的供给上限，迫使架构引入更多堆叠与横向扩展链路，从而加重布线密度与网络拥塞风险。

网络带宽瓶颈

当前，大模型训练通常依赖数千张 GPU 卡协同工作数周甚至数月，训练效率瓶颈并不仅仅取决于单 GPU 的算力，也受到 GPU 集群间通信效率的影响。GPU 间需进行频繁的梯度同步、参数更新、状态同步等集合通信操作，这些数据传递操作在服务器机内和机间均存在，且随着模型参数量的逐步提升，所传递的数据量也会不断增加。因此网络带宽越高，网络通信延迟在训练周期中占据的时间越短，也就能够提升 GPU 的利用率和有效计算时间占比。

以千亿参数规模的 AI 大模型为例，数据并行单次迭代的 AllReduce 集合通信数据量可达数百 GB 级别，如此庞大的数据量在极短的时间内需要完成传输与同步，对网络带宽提出了极高的要求。下表展示了不同模型规模单次梯度同步数据量的大小。

模型规模	典型 GPU 数量	单次梯度同步数据量	通信敏感度
十亿参数	数十卡	10GB 至 50GB	中等
千亿参数	数百至千卡	300GB 至 800GB	高
万亿参数	数千至万卡	大于 1TB	极高

电交换机的交换性能依赖其内部交换芯片，交换芯片的性能由其交换容量(switch capacity)衡量。交换芯片的交换容量 c 与交换机器的端口数 n、每端口的速率 r 满足关系：

$$c = n \times r$$

然而，由于交换容量 c 受限，交换机的端口数量 n 和带宽 r 无法同步提升，且其中一项的增加往往会导致另一项的下降，进一步制约了整体性能的提升。

网络延迟瓶颈

大模型训练需要多机多卡完成该轮所有集合通信操作后才可进行下一轮迭代，这种同步性特征要求智算网络必须提供极低的长尾时延，避免出现木桶效应[8]。根据理论推算，对于千亿参数规模的大模型训练来说，动态时延由 10us 增加至 1000us，GPU 有效计算时间占比将降低 10%左右。同样，大模型推理对网络时延也有着更高的要求，以确保能够为用户提供优质的推理服务。

传统数据中心网络普遍采用多层电交换架构，通过网卡与交换机连接多个计算节点，数据包在传输过程中需要经过多个交换节点的中转。受制于电交换“存储—转发”的工作机制，数据包必须在交换机内部进行排队等待，多层级的交换路径使得这一排队延迟被进一步放大，显著增加了端到端的网络时延，难以满足大模型训练对低延迟、高吞吐的严苛需求。

运行功耗瓶颈

大模型训练依赖数万张 GPU（如英伟达 A100/H100）并行计算。例如，GPT-3 训练耗电达 1,287 兆瓦时，相当于 3000 辆特斯拉跑 20 万英里的总耗电量。[9]根据 Lumentum 的研究，GPT-4 的训练网络功耗约为 21.5 MW；当扩展至 GPT-5（估计有 17.5 万亿参数、100,000 GPU）时，网络功耗飙升至 122 MW——超过 10% 的胡佛

大坝发电量。可见，电交换网络的功耗随训练规模呈指数级上升[10]。此外，大模型的训练周期长（如 GPT-4 需 90-100 天），GPU 持续高负载运行，且当前利用率仅 32%-36%，故障率较高，进一步延长训练时间并推高能耗。

为满足极端高速率转发需求，电交换芯片必须在高功率状态下运行，其高速 I/O 与大型转发能力意味着持续的高能耗（例如 CMOS 芯片在多层高负载下功率分布复杂）。与此同时，传统数据中心网络采用多层级的分层拓扑（如 fat-tree 或 Clos），在骨干和汇聚层中互连大量电交换元件，进一步扩大功耗基数与硬件复杂性。最终，这两方面叠加——单芯片高功率加上大规模设备自下而上扩展——使整体电交换网络的功耗急剧上升，在大规模模型训练与数据中心应用场景下具有供电不稳定的风险。

1.3.2 光交换的性能优势与发展趋势

光交换技术是智算中心网络架构的重大革新，其核心在于绕过了电交换芯片的物理限制，能够在极高速率下提供大量端口，极大地增强了智算中心的扩展能力。由于光信号特性，光交换采用线路交换机制，在发送与接收端之间建立专用光路，使数据能够以光速直接传输，从而大幅提升网络性能，光交换与电交换的区别如图 1-2 所示。

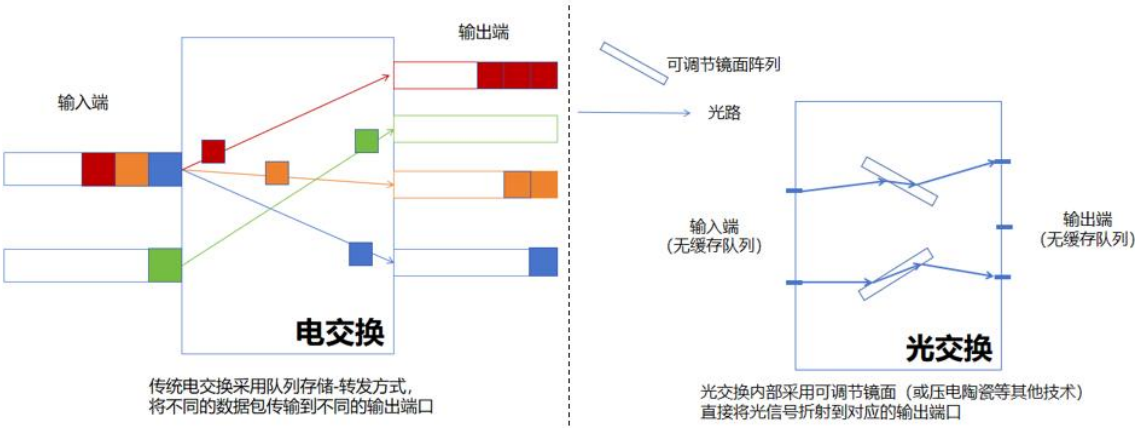

图 1-2 光交换与电交换原理对比

根据切换光路的执行方不同，主流光交换机可分为主动和被动两类，但主要的原理均为控制特定入射光的出射通路，借此“连接”入射链路和出射链路。

- **主动光交换机**通常利用 3D MEMS、液晶相位调制等原理，主要借助交换机内部元件的运动或物化性质改变等来改变光的出射方向。主动光交换机的重配置时间一般较长（数毫秒级），成本较高，但端口数量有明显优势，商用可达 320×320 个端口的规模，部分技术在实验室阶段已探索更大规模。
- **被动光交换机**典型的例子如 AWGR，则令不同波长的输入光在固定光路结构中被引导至不同的输出端口，从而实现波长选择性连接。AWGR 本身无任何可调组件，不具备动态重配置能力，其路径选择依赖于输入波长。系统级的路径切换可通过使用可调谐激光器来实现，切换速度取决于激光器性能，通常为微秒级。由于其通常无活动元件，所以其成本也较低，也无需电源；但端口数量通常较主动光交换机低，至高可达约 64~128 个端口。

尽管光交换技术具有高带宽、低延迟、可扩展等一系列优点，但在智算中心中应用全光交换面临诸多的现实挑战。首先，全光交换难以实现有效的缓冲机制，因为光信号无法像电信号那样轻松存储，这会导致在高负载的训练任务中，数据包冲突和丢包问题频发，影响任务的同步性和稳定性。其次，基于线路交换的特性使得光交换在灵活性和可重构性上受限，通常依赖于固定波长或空间切换，无法高效支持训练任务中频繁的动态通信模式，这可能导致网络瓶颈和资源利用率低下。现阶段使用光电协同方案组建智算中心网络，以结合光域的高速传输和电域的灵活控制，是更为实际的方案。

在传统的电交换数据中心网络中，常使用三层交换的网络拓扑，也就是网络的组织形式。最底层的叶交换机连接同一个机柜内的所有服务器，中间的聚合交换机连接多个叶交换机，而最顶层的骨干交换机则连接不同的聚合交换机。在光电协同网络中，通常根据需求不同，将光链路添加到已有电交换拓扑中，替换其上一层或两层。如图 1-3 所示。

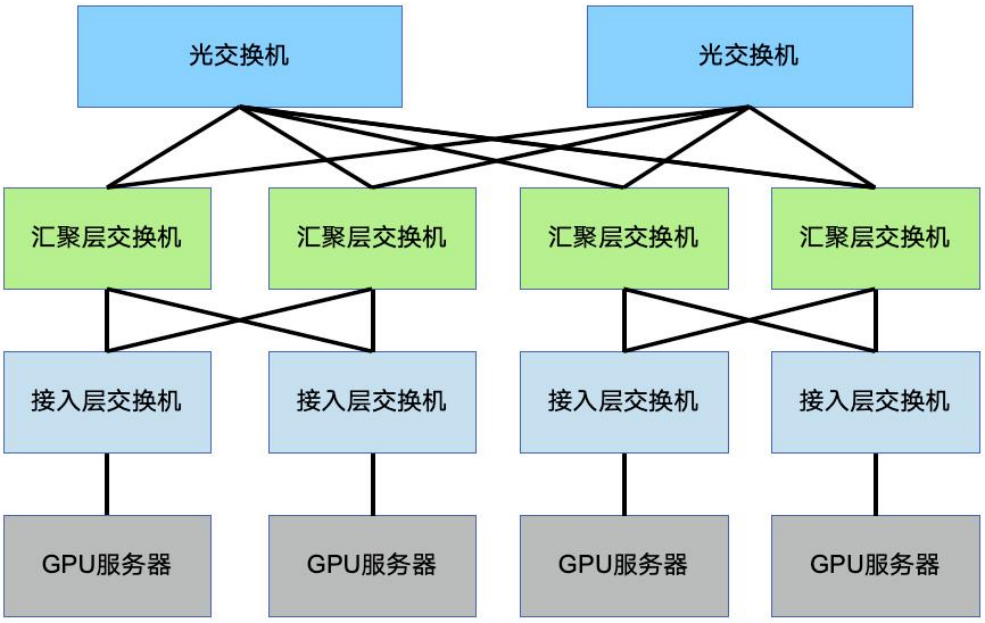

图 1-3 采用光交换机替代核心层的光电协同网络

光交换机制和光电协同网络相较于传统电交换网络，在多个关键维度上展现出显著优势，完美契合了智算时代对高性能网络的需求：

超高端口密度与扩展性：光交换技术通过创新设计突破了传统电交换在端口密度上的瓶颈。例如，基于 AWGR 相控阵干涉技术，可将数十至数百个波长解复用到单根输出光纤，而 MEMS 光交换机通过可旋转镜面结构，以线性成本增加端口，显著降低了扩展难度。超高端口密度可支持数千 GPU/算力节点互联，避免因端口不足导致的通信拥塞，保证梯度同步与参数服务器访问的高并发吞吐，为智算需求提供充足规模。以 400 GbE 为基准，一台 320×320 MEMS 光交换机能同时提供理论上无限的交换容量与 320 个 400G 端口，而要用 32 端口 400G 电交换机堆叠到同等端口数，常常需要 10 台以上，占用大量机柜空间，考虑拓扑则需同步增加汇聚、核心层交换机，进一步增加了布线复杂度。

无带宽瓶颈与未来演进性：光交换网络通过直接转发光信号，消

除了传统电交换中缓存读写及路由处理的速率限制。端到端光路的速率仅取决于发送端与接收端的光模块能力。更为重要的是，使用光链路时，网络速率升级仅需更换光模块，而无需替换光交换或光纤等核心部件。光交换的高速率且便于升级的特性，非常适合填补传统智算网络与需求的带宽差距。

低延迟与高效通信：光交换网络能够实现端到端连续光信号传输，这意味着数据在光域内直接传递，无需存储－转发，也避免了电交换网络中因排队、处理和多跳转发所带来的累积延迟。相比之下，电交换网络因为涉及光电互转（O/E-E/O）和芯片内部的存储－转发逻辑，其整体延迟大约在 30 μs 量级。大模型训练中进行 All-reduce 操作需要严格同步，或是对话式大模型进行实时推理时，这样的延迟差异会显著影响训练吞吐率与推理响应速度。

低能耗与可持续发展：光交换设备的能耗需求远低于传统电交换机。电交换设备功耗与比特率成正比，仅 32 口的 400GbE 电交换机就具有 420W 的典型功耗，峰值超过 1 kW；而光交换设备依据技术原理不同，如 320 端口 MEMS 交换机仅需 45W 典型功耗。同时，光交换技术减少了光-电-光转换的需求，进一步节省了光模块能耗。综合考虑来看，在一个使用 Fat-Tree 拓扑、具有 8000 块 GPU 和 400G 链路的数据中心中，仅将核心层 32 台电交换机替换为 9 台光交换机，可一并省下 2672 只 10W 功耗光模块，将核心层功耗由 62 kW 降低至 0.4 kW，节省逾 99%。大规模 AI 训练和推理集群往往成百上千机架并行运行，网络能耗占据数据中心总能耗相当比例。光交换的低功耗

特性不仅降低电力与散热成本，为 GPU 留出冗余，还为持续扩容的新一代算力平台提供绿色可持续的基础设施保障。

2. 智算中心光电协同交换网络面临挑战

在网络协议方面，智算中心网络通常遵循分层设计，与经典的 TCP/IP 五层模型一一对应：

- 应用层：面向大模型训练的集合通信操作（如 All-Reduce、All-to-All）；
- 传输层：RDMA（RoCE、IB）及定制化高性能通信协议；
- 网络层：路由控制与拓扑感知机制；
- 链路层：流量整形与无损以太网特性；
- 物理层：高速收发器、信号调制及光电转换技术。

然而，随着光电协同网络作为新一代数据中心架构的引入，传统分层协议栈面临新的适配挑战。这是因为光交换与电交换在基本通信模式上存在根本差异：电交换采用分组交换（packet switching），可灵活将同一链路上的连续数据包转发至不同目的地；而光交换采用线路交换（circuit switching），一旦建立电路，链路在拓扑周期内固定，无法并行服务其他节点，需依赖周期性拓扑重构实现全对全连接。

这一挑战异贯穿协议栈各层，促使现有的设计理念与机制需要重新审视与调整，以充分发挥光电协同架构的潜力。

2.1 应用层：集合通信与网络拓扑的失配挑战

智算应用中的集合通信操作（如 All-Reduce）通常抽象为特定的逻辑拓扑结构，包括树形、环形、蝶形等多种通信模式。每种逻辑拓扑都对应着不同的通信路径和数据交换顺序，其性能表现高度依赖于底层物理网络的连接特性。

当光电协同网络的物理拓扑配置与应用需求的逻辑拓扑结构不匹配时，会加剧热点链路的利用率问题，从而降低任务通信效率，同时造成其他链路空置：一方面，应用需要等待合适的直连链路建立，在等待期间造成带宽空闲；另一方面，链路建立后可能出现通信热点，部分链路过载而其他链路利用不足。

由于光交换机在任意时刻能够同时建立的链路数量有限，单个大规模集合通信操作可能无法获得充足的并行链路支持。这种资源约束要求在应用层实施更加智能的通信调度策略。

2.2 传输层：复杂功能的协议设计与流量调度挑战

传统电交换智算网络常常出现大量数据的传输需求，因此多使用高性能传输层协议（如远端内存直接访问 RDMA、定制化传输协议等），通过将数据流绕过处理器卸载到网卡，在相同处理开销下显著提升传输速率，与智算负载的高带宽要求非常契合。然而，智算中心对于互连网络的速度需求远超普通数据中心，而光网络的引入又带来网络核心带宽的进一步提升，因此对端侧网络协议栈的处理速度提出了极高要求。

当前，端侧高性能网卡的单端口速率已提升至 800 Gbps 量级，要求在极短时间内完成多个数据包的处理。这不仅对延迟控制提出极高要求，也使得片上存储难以直接支撑超高速处理流程。在连接数持续增长的背景下，分配给单条连接的协议栈资源进一步受限，高性能传输层协议在智算场景下面临三大核心挑战：

1. 多路径设计难题与乱序处理优化挑战

光电协同网络同时存在高速光链路、低速电链路以及暂时不可用的链路。传统基于哈希的多路径分配易导致流量落入非最优路径，而大规模维护链路状态信息又会引入显著开销。面向智算场景的传输协议需构建轻量级路径状态抽象，并支持快速、低成本的状态更新，以实现高效精确的负载均衡。光电协同网络在多路并发传输中容易出现数据包乱序。传统重排序机制依赖大量内存维护序列状态，难以适应片上存储空间紧张的高性能网卡。未来的协议栈需在有限资源下构建紧凑高效的数据结构与处理机制，实现对乱序数据的精准追踪与快速恢复。

2. 拥塞控制设计难题

传统传输层拥塞控制协议采用单态设计模式，主要针对同质化链路环境进行优化，通过统一的状态存储机制来管理多条相似路径，以降低协议的复杂度和存储开销。然而，光电协同网络中光链路与电链路在性能特征上存在数量级的差异：光链路通常具有数百 Gbps 的超高带宽和极低的延迟，而电链路的带宽相对较低但具有更好的延迟稳定性。这种巨大的性能差异使得单态拥塞控制协议在光电链路切换过

程中出现剧烈的速率和窗口震荡现象，严重影响协议的控制效果和网络稳定性。

3. 强潮汐流量的调度难题

不同于普通数据中心，智算中心的任务流量具有极强的潮汐性。大型智算任务的训练需要多轮迭代，每一轮迭代中都有本地计算和网络传输两个比较清晰可分的阶段。如何在传输开始时，迅速从间歇状态尽可能地快速用满网络带宽，是需要考虑的新问题。传统传输协议对底层网络拓扑变化不敏感，无法有效利用拓扑-流量的协同优化机会，进而根据链路状态选择合适的启动策略。如何在间歇状态不让网络资源空置，也需要全新的方式方法。

2.3 网络层：路由收敛滞后挑战

在传统数据中心环境中，网络拓扑结构相对稳定，边界网关协议(BGP)凭借其成熟的路由收敛机制能够有效应对偶发的拓扑变化，数十秒的收敛时间相对于静态的拓扑而言是可以接受的。

然而，光电协同交换网络的出现彻底改变了这一局面。光交换技术具备在数毫秒甚至微秒级别进行链路重配置的能力，拓扑切换频率较传统网络提升了数个数量级。在这种高频动态变化的网络环境中，依赖数秒间隔的Keepalive消息的传统BGP协议的收敛速度严重滞后于物理链路的重配置速度，导致路由信息与实际网络状态长期不一致，严重影响数据传输的可靠性和效率。

控制平面的扩展性挑战进一步加剧了路由控制的复杂性。传统

BGP 协议的收敛时间随着节点数量扩大而扩大，这与光电协同网络的扩展性优势背道而驰;每次重配置均需路由更新,重配置频率越高,要求单位时间内 BGP 收敛的次数便越高，给终端带来了远超传统数据中心 BGP 需求的性能开销，从而进一步影响智算任务运行效率。

近年来部分智算中心使用软件定义网络（SDN），以提供集中化、可编程的网络控制能力，但将传统 SDN 架构直接应用于光电协同网络时，面临显著的扩展性与时延瓶颈。集中式控制节点需要实时感知全网拓扑状态、计算最优拓扑配置，并将新的路由结果下发至各个网络设备。即便采用流水线化的计算与推送机制，整个流程的耗时依然可能达到秒级,远无法满足光链路数十微秒至百纳秒级的拓扑更新需求。随着网络规模的不断扩大，单一控制节点还会成为性能瓶颈，进一步延长路由配置下发的时延。

这类时间尺度的不匹配，使得路由信息在光链路拓扑切换后可能长期滞后于真实网络状态,导致数据包被转发至失效路径或次优路径,从而引发链路中断、丢包增加及延迟上升等问题。在光电协同网络中,拓扑切换频繁,传统依赖秒级收敛的路由协议与集中式控制架构难以适应这种高动态环境。

2.4 链路层：非对称资源动态分配挑战

智算中心长期面临链路带宽利用不均衡的问题。以典型的智算训练任务为例,参数推送与参数拉取之间的流量存在数量级差异。传统电交换网络多采用固定对称的全双工链路配置,以适应多变的负载需

求并降低人工调整成本。然而，这种静态对称配置模式下，当上行链路达到饱和时，下行链路常常仅用于传输少量确认应答（ACK）等控制信息，带宽资源未能被充分利用，导致严重的带宽浪费。

相较于电链路，光链路天然具备动态重配置能力，能够灵活调整节点间链路的分配比例，支持针对不同节点对的上下行链路进行独立配置，甚至实现单工模式传输。这种灵活性为解决链路带宽利用率不均提供了技术基础。

然而，目前传统的光电协同网络部署仍多沿用对称的上下行链路配对设计，虽保障了全双工传输的能力，却未能根本打破上下行带宽固定对称的局限性，无法充分挖掘光链路动态重配置的优势，带宽资源浪费问题依然突出。

2.5 物理层：信号衰减挑战与时延约束挑战

在智算中心网络中，物理层不仅承担着数据的信号传输职责，更直接影响整个系统的带宽、时延和稳定性。然而，当光交换技术被引入到智算网络时，物理层面临一系列前所未有的挑战。

首先，光互连需要在高端口密度和长距离传输之间取得平衡。相比传统电互连，光链路在传输距离上具有天然优势，但这也意味着必须控制光信号衰减和插损，否则网络性能将随着互连规模的扩大而显著下降。当前业界方案普遍依赖硅光模块和高速收发器，但在密集布线场景下，插损累积、反射干扰以及光纤管理仍是制约因素。

其次，光交换的动态性带来了更高的控制复杂度。与固定电交换

架构不同，光电协同网络需要根据训练任务和流量模式动态切换光路。然而，光器件的切换速度和稳定性远未达到纳秒级，这使得在多租户、混合负载环境中保持低延迟成为重大挑战。特别是当光交换元件分布于多个 GPU 节点时，其控制信号同步、状态感知及故障恢复需要更复杂的协议和硬件支持。

再者，物理层还需应对能效和散热问题。高速光模块和大规模硅光阵列的能耗不容忽视，尤其是在分布式光交换架构中，交换功能被集成至 GPU 互连模块或板卡，这对光器件的功率预算、热管理提出了极高要求。此外，工艺制程的限制也是一大瓶颈。尽管硅光技术降低了对先进半导体制程的依赖，但其集成度、良率和一致性仍需长期优化。由此可见，光电协同架构在物理层面不仅面临信号完整性、动态可重构性和能耗管理的综合挑战，还需解决制造与生态成熟度不足的问题。

3. 智算中心光电协同交换网络协议栈技术发展

在第 2 章中，本文已对智算中心网络在采用光电协同交换架构时所面临的软件协议栈技术挑战进行了系统分析，涵盖了应用层、传输层、网络层、链路层及物理层等多个层面。针对上述挑战，本章将提出一些具有代表性的技术方案和解决方向，旨在为相关研究提供思路启发。需要说明的是，这些方案并非唯一或最优解决方案，而是从多种可能的技术路径中选取的典型方案，以期抛砖引玉，促进该领域的

进一步探索与创新。本章的整体结构如图 3-1 所示。

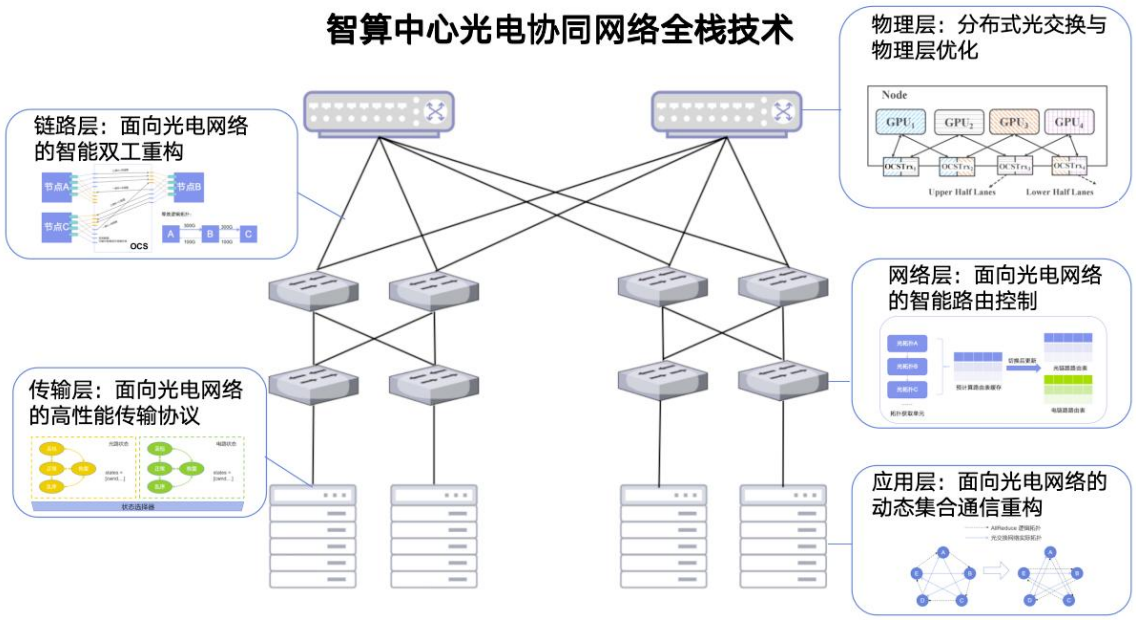

图 3-1 智算中心光电协同协议栈技术

3.1 应用层：面向光电网络的集合通信重构协议

在智算中心的运行环境中，分布式机器学习训练构成了最核心的计算负载。为了在成百上千个处理器(如 GPU)之间实现高效的数据交换和同步，相当大比例的通信任务采用结构化的集合通信原语形式进行组织。典型的集合通信操作包括 AllReduce、All-to-All、Broadcast 等，这些操作在实际训练过程中的通信开销可占据整个训练时间的 20%至 50%[11]。

集合通信拓扑失配问题是光电协同网络面临的核心挑战之一。由于光交换网络的稀疏连接特性，传统为全连通网络设计的集合通信算法在光电协同环境中经常出现拓扑不匹配现象。例如，标准的 Ring AllReduce 算法假设所有节点能够形成完整的环形连接，但在光交换

网络中，由于物理光链路数量限制，仅有部分节点对能够建立直连光链路，导致传统环形拓扑无法有效映射到实际物理连接上，从而产生通信阻塞和性能严重下降。

强潮汐流量的调度优化是另一个关键问题。智算训练任务呈现出明显的潮汐特征：在模型训练的不同阶段，通信负载的时空分布极不均匀。传统的静态资源分配策略无法适应这种剧烈的负载波动，导致在高峰期出现严重的网络拥塞，而在低谷期大量带宽资源被浪费。

3.1.1 预测通信模式，为重配置提供需求启示

针对光电协同网络的动态重配置特性，应用层需要具备前瞻性的通信模式预测能力，为物理层的链路重配置提供精准的需求指导。机器学习负载中的许多通信模式具有高度的可预测性特征。例如，在专家混合(MoE)模型的专家路由阶段，系统会产生大规模的 All-to-All 通信需求；而在数据并行训练的梯度同步阶段，则主要表现为大容量的 AllReduce 操作模式[12]。

基于这一发现，可以在应用层构建智能的通信模式预测机制。该机制根据当前模型架构特征、训练阶段状态以及历史通信行为，主动预测即将出现的数据通信需求模式。这种预测能力使系统能够提前调整光网络的链路配置，优先为关键集合通信操作建立直连链路，并通过重配置时间的预先调度来最小化链路切换等待延迟，实现通信性能的显著提升。

3.1.2 拓扑有感知的动态集合通信重构

现有的机器学习训练框架中，开发者通常直接调用通信库(如NVIDIA NCCL)提供的标准集合通信算法实现。然而，这些通信库往往只内置了有限的几种算法变体，例如 NCCL 中的 AllReduce 操作默认采用 Ring AllReduce 算法实现。在光电协同网络环境中，由于拓扑结构的动态性，固定的环形连接方式经常无法与实际的物理链路配置完全对应，导致部分关键链路缺失和通信过程阻塞。

要实现集合通信操作与网络拓扑的深度适配，需要在通信库中为每种集合通信原语内置多样化的算法实现方案，包括树形、环形、分层混合等不同的拓扑组织模式，并构建智能的算法选择机制，根据当前网络拓扑特征动态选取最优的通信算法。例如，当 GPU 集群以树形拓扑方式连接时，采用树形 AllReduce 算法能够获得明显优于环形 AllReduce 的性能表现；针对稀疏光连接设计，可以采用分段环形 AllReduce，将全局环形分解为多个子环形，每个子环形内部使用电链路，子环形之间通过光链路连接；还可以进一步结合数据并行和模型并行的特点，使用不同的 AllReduce 实现。

以图 3-2 中的拓扑为例：五台服务器执行 Allreduce 通信。采取固定的 Ring Allreduce，每条流会被中转两次（如服务器 A 到服务器 B 的流需要经过 C、E，才能被转发到 B），等效带宽变为链路带宽的 1/3，网络时延也因为在中间节点的服务器上排队而大幅增加。如果讲 Allreduce 的通信顺序与当前光网络拓扑匹配，则每个节点都能最大化利用链路带宽，且没有额外的网络延时。

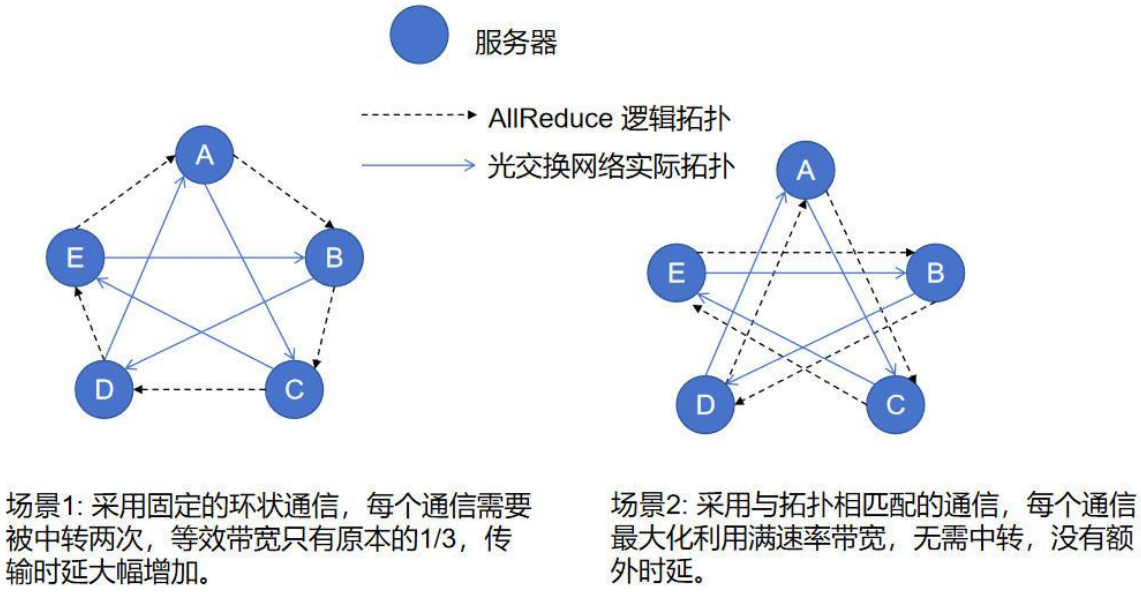

图 3-2 采用与拓扑相匹配的集合通信

实际实现上述技术时，需要考虑许多问题：

控制平面的复杂度问题：在实际部署过程中，控制平面需要同时处理来自大量并行训练任务的信息收集、分析和决策制定工作。系统必须从海量的任务执行数据中快速提取通信模式特征，并计算出最优或接近最优的网络拓扑配置方案，随后协调各个网络组件完成重配置操作。为了应对这种复杂度挑战，可以采用分层控制架构，将全局优化决策与局部快速响应相结合，通过边缘智能和预计算技术降低实时决策的计算负担。

重配置时延问题：对于采用 MEMS 技术的光交换设备，数毫秒级别的重配置开销对于许多时延敏感的通信任务来说仍然过高。为了进一步提升光电协同网络的性能上限，需要推动更加快速的光交换技术发展，同时在系统层面通过预测性重配置、并行重配置和重配置与通信的流水线处理等技术手段来掩盖和减少重配置延迟的影响。

标准与可交互性：实现通信模式动态重构需要对底层通信库进行

深度修改或开发全新的通信库实现。如何在提供先进的通信模式重构能力的同时，保持与现有代码生态的良好兼容性，成为了技术落地过程中必须解决的关键问题。建议采用逐步演进的策略，通过 API 层面的抽象，使新技术能够无缝集成到现有的机器学习框架中，降低用户的迁移成本和学习门槛。

通过上述技术方案的系统性实施，通信模式动态重构技术能够显著提升光电协同网络在智算中心应用场景下的性能表现，为下一代高性能计算基础设施的构建提供重要的技术支撑。

3.2 传输层：面向光电网络的高性能传输协议

高性能传输层协议作为现代数据中心的核心通信技术，通过绕过处理器、减少数据拷贝等机制，能够显著降低 CPU 占用率并提升数据传输效率。然而，这类协议的设计与优化大多基于全电交换网络的特性，当面对光电协同网络的独特环境时，在多路径传输、拥塞控制、流量调度以及数据完整性保障等关键方面，均暴露出明显的适配性不足问题。

在光电协同网络中，光链路与电链路在带宽容量、传输延迟以及连接稳定性等方面存在显著差异，传统高性能传输层协议的单态设计理念已无法有效应对这种高度异构的网络特性。为了充分释放光电协同网络的潜力，需要对传输层协议栈进行系统性重构与优化，使其能够感知并动态适配不同链路类型的性能特征，从而实现跨介质的高效、可靠数据传输。

3.2.1 灵活的多路径传输机制

在传统的电交换智算中心网络中，多路径机制已相对成熟。这些机制通常建立在网络层的等价多路径（ECMP）路由基础之上，并进一步扩展出多路径感知的拥塞控制、更智能的路径选择，以及更高效的乱序重排等高级特性。然而，在光电协同网络环境中，这些机制面临全新的挑战，包括路径可用性的高频动态变化、传输过程中的流迁移、以及不同链路间延迟差异导致的乱序处理问题。

为适应光电协同网络的动态特性，需要构建具备实时路径感知与决策能力的智能多路径机制。传输层协议必须与控制平面紧密协作，及时获取当前可用链路状态，并结合光网络的重配置计划以及应用层的通信模式预测，对不同流量特征进行分类处理，提前规划最优传输路径。在实际运行中，当某条流量已在电链路上传输时，迁移至光链路可能带来显著性能提升，也可能得不偿失。因此，在执行迁移决策前，需要结合链路负载情况与潜在收益进行评估，并通过虚拟链路与物理链路的映射机制，避免切换路径导致的哈希五元组改变，从而减少 ECMP 重新分配带来的不可控影响，如图 3-3 所示。

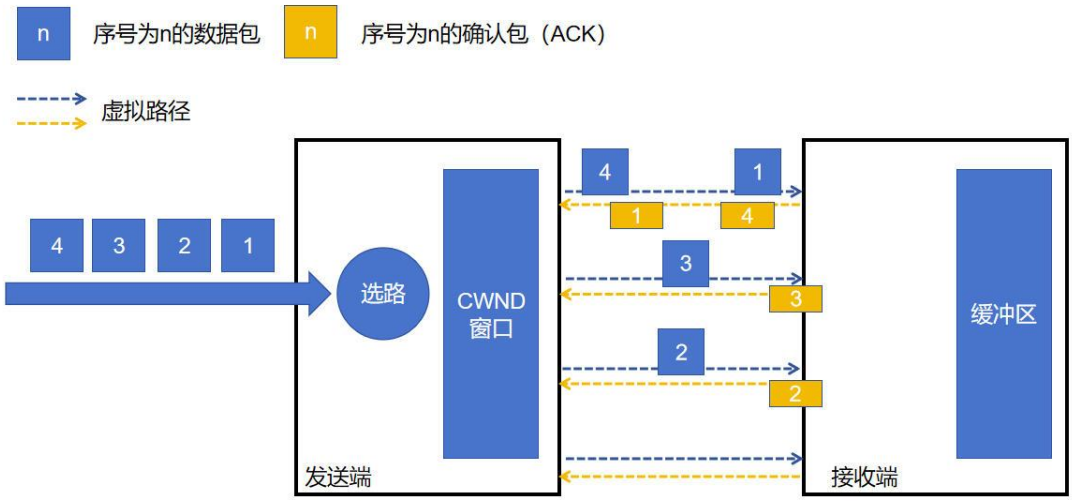

图 3-3 采用虚拟路径的多路径传输

光路与电路在传输延迟上的差异，还会导致一种典型现象：先经电路发送的数据包，可能晚于后续经光路发送的数据包到达接收端，从而引发乱序。传统传输层协议在面对乱序时，往往选择直接丢弃并重传，这种处理方式虽然简单，但会严重削弱光电链路切换的效率。应对这一问题有两类潜在优化方向：其一是在发送端对光链路的发送进行精确延迟，使其在启动传输时，电路上的先发数据已接近到达接收端，从而降低乱序发生的概率。这种方法在牺牲少量光链路吞吐的前提下，可以保持数据到达顺序的稳定性。其二是在接收端引入增强型乱序重排机制，将来自光路的乱序数据临时缓存，待电路数据到齐后再进行有序交付。这种方法能显著减少链路空闲时间，提升整体吞吐，但需要接收端具备更高的硬件能力，尤其是在 DRAM 容量与数据查找速度方面，以应对数十毫秒级的高吞吐缓存需求。

3.2.2 双状态拥塞控制机制

在传统的电交换网络中，传输层的拥塞控制协议通常采用单态设

计模式，通过统一的状态存储机制来管理多条特性相似的路径，以降低协议复杂度和状态维护开销。然而，在光电协同网络中，光链路与电链路在性能特征上存在数量级的差异：光链路往往具备数百 Gbps 的带宽和亚微秒级延迟，而电链路的带宽相对较低但延迟更加稳定。

这种显著的性能异构性，使得单态拥塞控制在光电链路切换过程中容易出现传输速率和发送窗口的剧烈震荡，削弱了控制的稳定性与收敛速度。为此，需要引入光电双态拥塞控制机制，为光链路与电链路分别维护独立的发送窗口、速率参数及控制逻辑，以实现针对性优化,如图 3-4 所示。

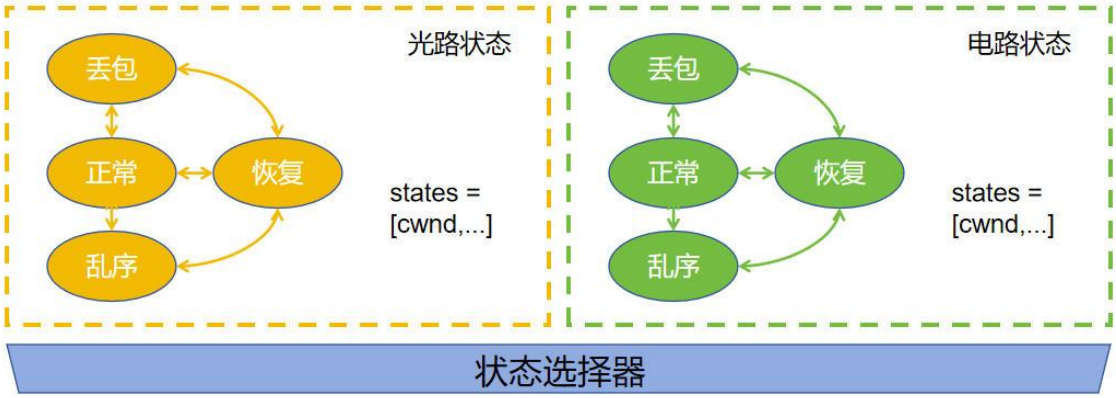

图 3-4 双状态拥塞控制

双态机制的核心挑战在于如何实现拥塞控制状态与链路物理状态的精确同步。目前可行的检测方案主要有两类：

显式信号通知机制：在链路状态切换时，通过网络控制平面发送专用信令，显式告知发送端切换至对应链路模式。例如，当光链路建立时，触发光链路模式并更新相关控制参数；在即将进行拓扑重配置前，再切换回电链路模式。这一方式同步精度高，但会引入额外的信令开销。

主动探测机制：由发送端通过定期探测流判断链路状态变化。当检测到光链路可用时，主动切换至光链路模式；光链路释放或不可用时，再回退至电链路模式。这种方式减少了信令依赖，但切换时机可能因探测延迟而滞后。

此外，在传统网络环境中，多流竞争常采用公平带宽分配策略，以确保各流获得均等的传输机会。然而，在智算中心的分布式训练场景下，公平分配并不一定最优。研究表明，在计算量固定的条件下，训练任务的总体完成时间更多取决于通信阶段(如参数同步)的耗时。因此，非公平的流量调度策略可能更高效：通过优先满足部分任务的通信带宽需求，使其尽快完成传输，从而将不同任务的通信阶段与计算阶段交错安排。在理想情况下，通信负载可以被均匀分布在时间轴上，实现全生命周期内带宽资源的高效利用。

3.2.3 错峰出行智算流量调度方案

非公平调度在智算中心的核心目标并非追求带宽“公平”，而是通过有意识地错开多个训练任务的通信窗口与计算窗口，降低并发通信的峰值负载，从而提升总体资源利用率与任务完成效率。具体思路是将每个训练任务的生命周期视作交替出现的“计算阶段—通信阶段”序列，调度器有选择地安排某些任务保留网络资源，使其他任务在其计算阶段完成时占用网络进行通信。通过这种非公平分配，可以显著减少同时处于通信阶段的任务数，平滑带宽需求波动，缓解瞬时拥塞并降低尾部传输延迟，如图 3-5 所示。

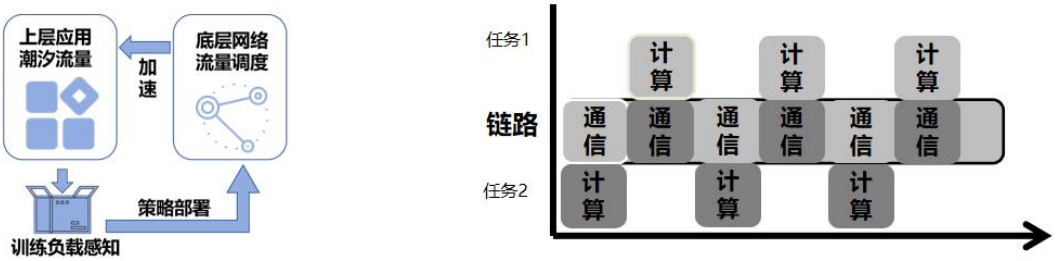

图 3-5 错峰出行的流量调度策略

要实现上述策略，需要三方面能力的协同：一是准确的阶段预测与标注（来自作业调度器或应用层的剖面信息），以便识别各任务的通信高峰；二是时序化的带宽资源管理能力，支持为被选定的任务预留或独占链路时隙，并能在光链路重配置窗口内快速生效；三是跨层的控制接口，使应用层、作业调度器、网络控制平面与传输协议能够共享时序信息并执行协调决策。实现上可采用基于时间片的“通信抢占/授权”机制、任务优先级队列与短期资源预占，以及必要的回退策略（当预测失准或链路不可用时的迅速回退到公平模式）。

该策略带来的直接收益包括：显著缩短单个训练任务的通信关键路径时间、提高光链路的有效利用率、降低整体集群的带宽争用风险，并通过错峰传输提高系统吞吐与作业完成率。但同时也带来公平性与调度复杂度的权衡——非公平分配需要合理的策略和可观察性保障，以避免长期性资源饥饿或调度振荡。因此，在实践中应以全生命周期完成时间（job completion time）和系统吞吐为主要优化目标，辅以可配置的公平性约束和动态调整机制，逐步在生产环境中验证与推广。

3.3 网络层：面向光电网络的智能路由控制

网络层作为网络协议栈的核心组成部分，包含控制平面和数据平

面两个关键子系统。数据平面专注于高速数据报文的转发处理，而控制平面则负责路由决策、拓扑发现和转发表维护等智能控制功能。

3.3.1 路由协议的光电优化方向

传统 BGP 协议具有诸多缺点，已经不能适应光电混合只算中心网络的发展：收敛过慢导致路由信息长期滞后于实际网络状态，大规模收敛开销大导致难以扩展网络规模。为了令 BGP 协议适应光电协同网络场景，一种思路是对其进行简化，以便降低开销并提高收敛速度，几种可能的简化方向如下。

减少无用的属性处理。标准的 BGP 协议包含大量针对广域网环境设计的属性和机制，而其在数据中心环境中用途有限。可以考虑将冗余属性移除或进一步整合，同时采用更直接的路径选择方式，以降低 CPU 和内存开销，为快速收敛提供有利条件。

调整链路探测方式。传统 BGP 协议采用基于定期 Keepalive 消息的链路探测机制，这种被动式的状态发现方法在光电协同网络的快速变化环境中显得过于迟缓。通过根据网络动态性的不同阶段智能调整探测间隔，或是使用主动的链路变化反馈，实现事件驱动通知，显式推送链路变化状态，并融合应用层流量统计、物理层信号强度、链路层状态等多种信息源，BGP 协议可以更早地探测到链路变化，降低甚至避免被动探测带来的时延。

换用其他传输协议。BGP 协议基于 TCP，其建立连接需要复杂的握手机制，且复杂的拥塞控制和重传机制也对 TCP 引入了额外的开

销，不仅增加了通信延迟，而且对频繁的 BGP 协商不利。通过换用 UDP 协议或 RDMA 协议，可以实现更低延迟的传输。

预留专用控制通道。在光电协同网络中预留专用的控制平面链路，可以在保证全对全连接的情况下免受数据平面流量排队影响，获得稳定的传输性能。

3.3.2 面向光电拓扑的预计算优化与双模路由表设计

由于分布式机器学习的流量周期性，光电协同网络常常使用数种固定的拓扑，并按固定的时间间隔进行轮转。利用这一特性，可以进一步提前将每种拓扑对应的路由决定缓存在各节点内，从而免去了光路重配置时所需发送的完整下一跳信息。

在指定轮转规则前，可以针对数种典型拓扑模式，预先计算其可能的路由表配置，使其在网卡缓存的承受范围内；当网络需要切换到特定拓扑时，可以免去 SDN 计算或 BGP 收敛延迟，提早完成路由的建立，如图 3-6 所示。

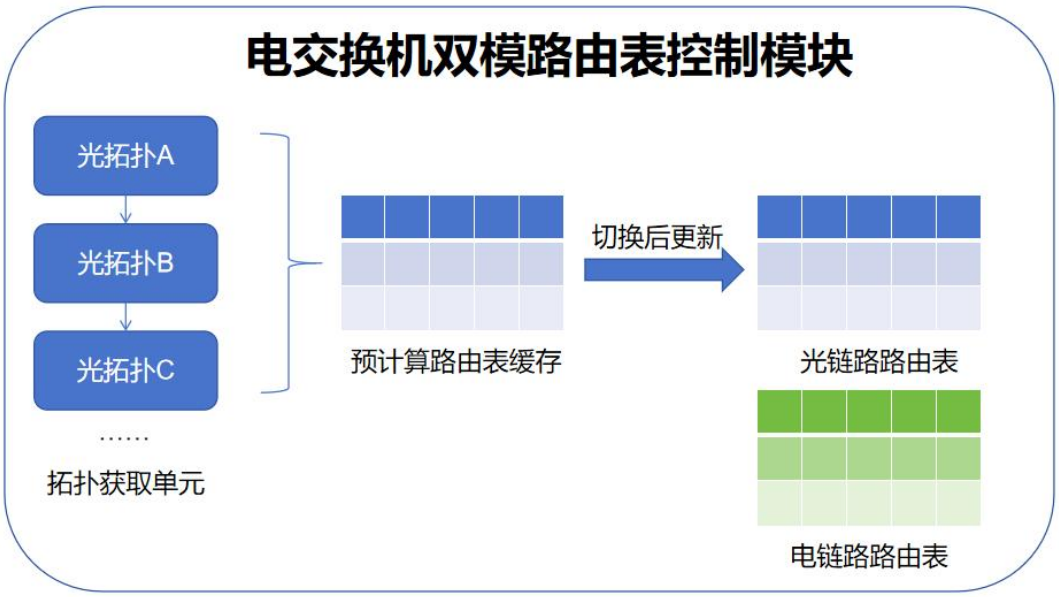

图 3-6 双模路由表设计

对于部分光电协同网络拓扑类型，在特定时间可能同时存在电链路和光链路，这两种链路的显著差异决定了其适合不同类型流量。一方面，可以设计双模路由表，为光链路和电链路维护独立的路由策略，另一方面，根据链路带宽、延迟、可靠性等综合评估链路，计算各自的优先级，并调度流量。最后，当光链路出现故障时，应用故障切换方案，切换到电链路备份路径，保证服务的连续性。

通过上述技术方案的系统性实施，智能路由控制体系能够显著提升光电协同网络的路由收敛速度和整体性能，为智算中心提供高效、可靠的网络层支撑，推动下一代数据中心网络技术的发展和应用。

3.4 链路层：面向光电网络的智能双工重构

智算中心普遍存在带宽资源配置失衡的核心问题。在智算训练场景中，数据上传与下载的流量需求呈现极不对称的分布特征，两者间往往存在数十倍甚至百倍的差异。传统电交换架构受限于静态配置模式，通常采用固定的双向等带宽设计，难以响应动态变化的业务需求。这种刚性配置导致网络资源配置与实际使用需求严重脱节——当某一方向链路负载接近满载时，反向链路的带宽利用率可能极低，造成大量网络资源闲置。

光交换技术为解决此类问题提供了新的可能性。其核心优势在于支持灵活的连接重构：既可以在宏观层面重新映射节点间的连接关系，也可以在微观层面独立调整单一节点对的上下行通道配置。但传统的

配对式双向链路部署方式，虽然确保了通信的双向性，却延续了带宽均等分配的局限性，未能从根本上缓解资源配置与需求不匹配的矛盾。

3.4.1 上下行非对称带宽的链路利用

智算中心的典型应用负载呈现出显著的流量方向性特征。如数据并行的 AllReduce 通信，流水线并行的点对点 Send/Recv 通信，这些集合通信的部分方向存在显著的大流量，而反方向只需要返回确认 ACK 数据包,如图 3-7 所示。

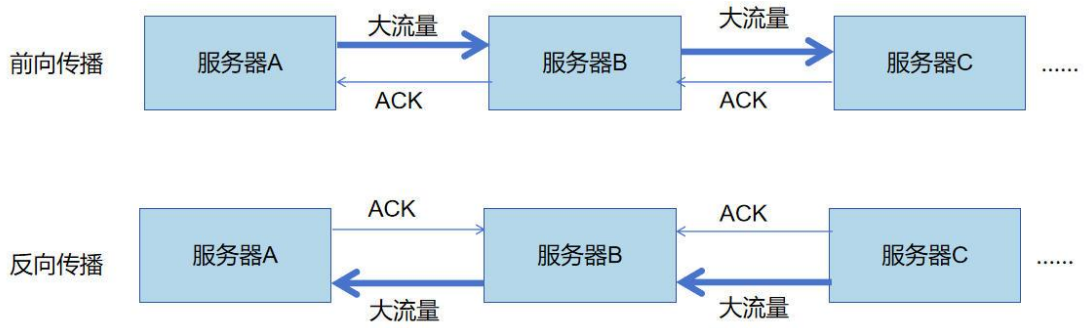

图 3-7 流水线并行中的流量不对称关系

在联邦学习场景中，边缘节点向中央聚合器上传本地模型更新，而中央服务器向各节点分发全局模型，通常上行聚合流量大于下行分发流量。在推理服务场景中，客户端向推理服务器提交相对较小的查询请求，而服务器需要返回大容量的推理结果，形成明显的下行流量占优模式。对于对称全双工链路，在前一个例子中，工作节点到参数服务器的链路带宽被浪费；而后一个例子中，向各个边缘节点方向的链路利用率较低。

3.4.2 智能预测与链路池化资源管理策略

基于以上的发现，根据实时流量需求的方向性特征，利用可灵活重构的光链路调配流量需求成为了一种解决思路。为了更好地匹配流量需求与链路带宽，需要准确预测未来时段的流量需求模式，通过跨层设计克服链路层的局限性。首先，通过分析上层应用的执行阶段和通信模式，可以预测即将到来的流量方向性。如在参数同步阶段，可能出现大量的聚合通信需求。其次，历史流量特征可能表露出周期性与趋势性，以此为未来需求变化提供参考，影响带宽分配决策。最后，须在网络中部署细粒度的流量监测机制，实时感知上述两种机制未能预测的流量需求，并及时调整带宽分配策略。

在具备预测能力的基础上，链路池化策略成为提升带宽利用率的核心。通过光交换架构实现端口输入与输出的解耦，系统能够将多个物理光纤端口抽象为一个共享带宽池，并通过动态配置实现不对称带宽分配。例如，当预测到 A→B 的数据流量在下一个调度窗口显著高于 B→A 时，调度器可以在光交换层面建立多条 A→B 的光路，形成 300G 对 100G 的非对称带宽映射，如图 3-8 所示。从而充分利用节点可用的发送端口资源。

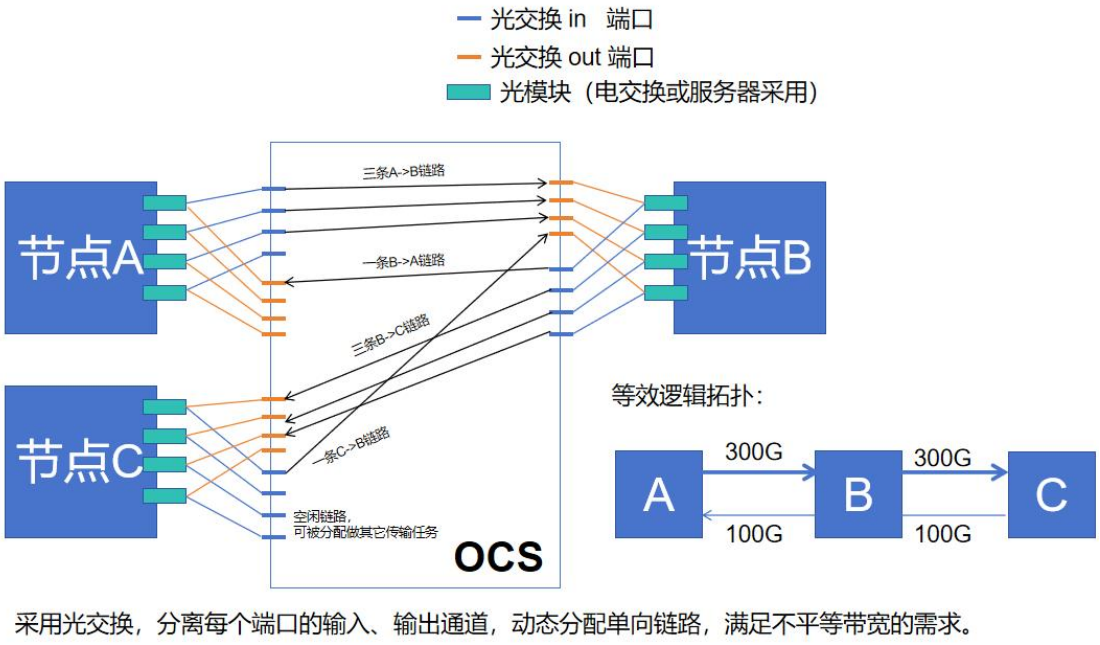

图 3-8 光交换动态配置实现不对称带宽分配

为了最大化带宽资源的利用效率，可以将光链路的分配抽象为资源池化的模式进行管理。可以将物理光链路抽象为虚拟链路资源池，每个虚拟链路可以根据需求动态调整带宽容量和方向属性。在全局范围内进行链路资源的监测与优化调度，当某些区域的链路资源出现空闲时，可以将这些资源重新分配给其他有更高需求的区域使用。根据流量负载的实时变化，可以通过增加或减少物理光链路等方法，动态扩展或收缩虚拟链路的带宽容量，实现资源配置与需求变化的实时匹配。

要做到不对等的带宽分配，链路层需要对现有协议和机制进行扩展，主要包括以下几个方面：

1. 链路抽象与逻辑接口增强

传统链路层假设物理链路带宽固定且对称，因此需要引入“虚拟链路”抽象，允许将多个物理通道聚合为一个逻辑带宽池，并通过链路

层管理平面向上层提供动态带宽分配能力。

2. 链路状态信令扩展

标准的链路层协议（如以太网 LACP、OTN 控制信令）通常不携带瞬时带宽配置的信息。为支持动态调整，需要扩展链路状态信令，增加可变带宽描述字段，使控制平面能够快速通知各端节点当前有效带宽，避免拥塞或速率不匹配。

3. 速率自适应与流量整形

链路层需支持根据调度器下发的带宽分配策略，对发送速率进行自适应控制。例如，在 A→B 分配 3 条通道、B→A 分配 1 条通道的场景中，节点 A 的 NIC 必须限制反向发送速率，同时节点 B 必须感知该配置，避免过量发送导致丢包。

4. 与上层路由/传输的协同接口

链路层需要向上层暴露动态带宽变化事件，以触发流量调度或路径选择优化，防止因链路调整造成路由震荡或传输层协议拥塞控制的性能波动。

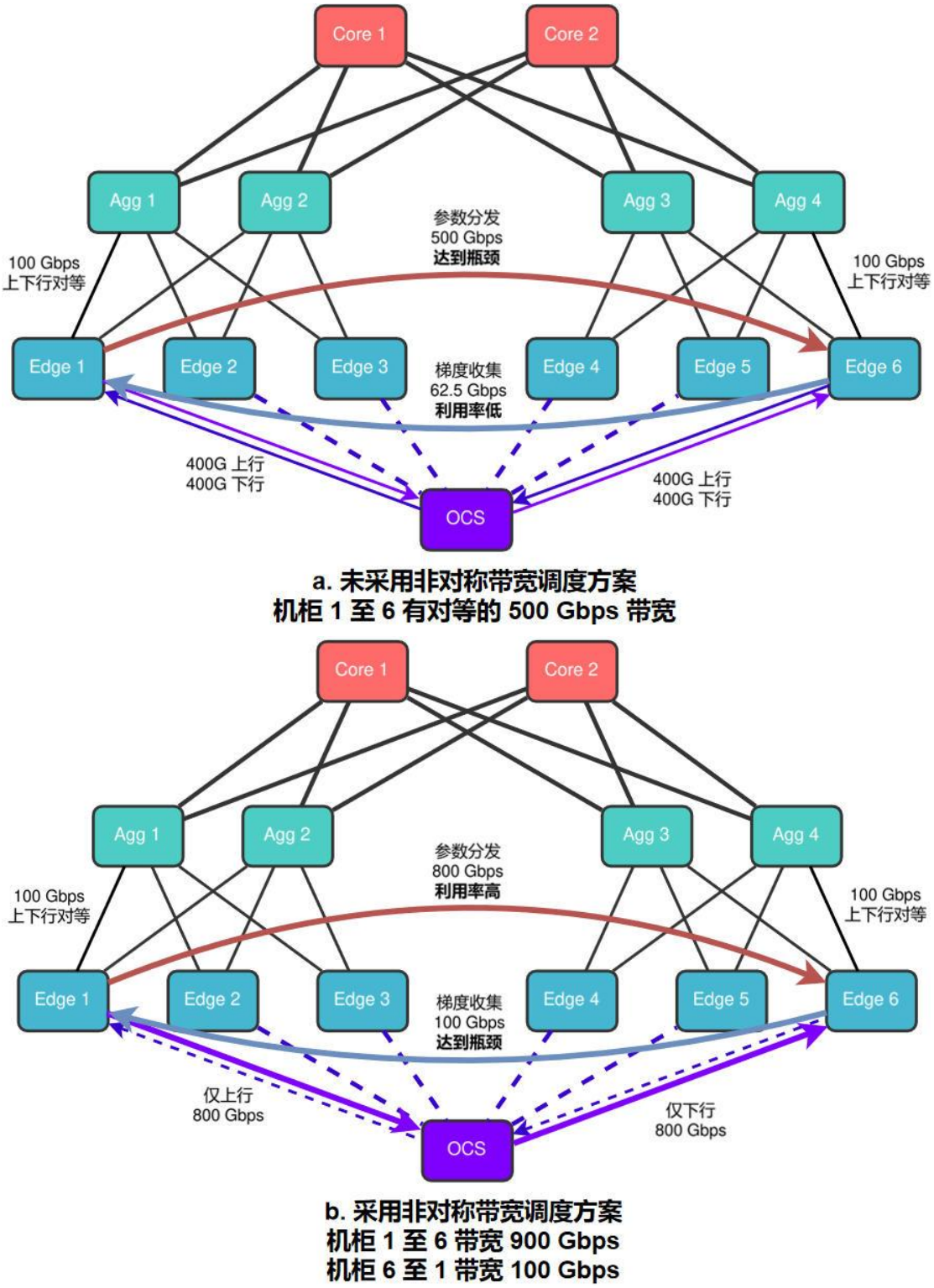

图 3-9 链路层采用非对称带宽方案对比

下面通过理论计算，估计智能双工重构带来的技术提升。如图 3-9 所示，对于一个使用光链路增强传统 Fat-Tree 拓扑的智算集群，在电网络基础上额外使用若干光交换机连接 ToR，光网络带宽 400 Gbps，

聚合层及以上的电网络带宽 100 Gbps，负载为分布式机器学习训练任务，参数分发与梯度收集的流量比设为 8:1。在传统对称配置中，除了现有电链路之外，在发送机柜 1 和接收机柜 6 之间分别建立一条上下行方向光链路，则上述两个 ToR 之间聚合上下行带宽均为 500 Gbps，包含 100 Gbps 的电链路和 400 Gbps 的光链路。令参数分发负载以 500 Gbps 运行，则梯度收集以 62.5Gbps 运行。上述电链路和两条光链路的总体利用率为 56.25%，而后者链路的利用率仅 12.5%。应用智能双工重构后，将两条光链路均应用于参数分发方向，此时上下行链路带宽分别为 900 Gbps 和 100 Gbps，其中上行包括两条 400 Gbps 的光链路和一条 100 Gbps 的电链路，下行则仅包括一条 100 Gbps 的电链路，当下行链路达到瓶颈时，参数分发和梯度收集吞吐量分别为 800 Gbps 和 100 Gbps，此时总体利用率为 90%，相比应用前有了 60%的提升。

通过上述技术方案的系统性实施，智能双工重构技术能够有效解决光电协同网络中的带宽利用不均衡问题，显著提升网络资源的利用效率，为智算中心提供更加灵活高效的链路层支撑，推动光电协同网络技术在实际生产环境中的成功应用。

3.5 物理层：分布式光交换与物理层优化

业界正在探索多种物理层优化技术，以推动光电协同网络加速落地。其中，分布式光交换（dOCS）架构作为一种代表性技术方案，受到了广泛关注。该架构的核心思路是将光交换能力前移至 GPU 节点，通过在每个 GPU 互连模块内嵌光交换功能，实现网络拓扑的弹

性可重构。与集中式光交换方案相比，这种分布式设计有效缓解了光纤汇聚带来的布线复杂度，并可通过控制平面实现秒级拓扑重构，显著提升网络的灵活性与可靠性。曦智科技的 LightSphere X 超节点[13]是该技术路径的典型实现之一。

在器件设计层面，硅光技术是另一个重要发展方向。该技术将调制器、开关阵列与光电转换功能集成至 Chiplet 模块，并可结合先进封装技术缩短光电互连路径，降低损耗并优化信号完整性。通过超低损耗载板和多层互联架构，此类方案在实现高端口密度的同时，能够保证物理链路的稳定性与低误码率。光模块的散热优化设计也是关键技术要素，确保系统在高带宽、长时间训练任务下保持稳定运行。

为了应对光交换切换速度的物理瓶颈，行业方案引入了多级控制策略和预测性调度机制，将拓扑切换与训练任务调度协同规划，减少全局重构的频率，从而降低光切换带来的延迟波动。与此同时，未来的物理层优化方向还包括采用 CPO（Co-Packaged Optics）实现光模块与交换芯片的深度集成，以及探索液晶可调开关、MEMS 等新型光开关器件，以在降低功耗的同时进一步提升切换速度。随着这些技术的演进，光电协同网络在物理层有望实现低延迟、高扩展性与能效优化的平衡，为智算中心迈向万卡规模互连奠定坚实基础。

4. 总结与展望

综上所述，第 1 章分析了智算中心在政策、业务和技术层面的发展趋势，并揭示了智算中心光电协同网络的兴起，第 2 章详细分析了

光电协同技术在智算中心面临的技术挑战，第 3 章提出了光电协同智算中心网络协议栈的技术发展策略。在此基础上，本章将总结光电协同网络的发展路径，提出标准化路线图，并展望其未来技术演进与应用前景。

4.1 光电协同交换网络的标准化路径

光电协同网络作为下一代数据中心网络架构的重要发展方向，其标准化工作关系到整个产业生态的健康发展。基于前述四层技术方案的深入分析，本路线图从技术标准、产业协作、实施部署和生态建设四个维度，按照网络分层从上到下的顺序，制定了分阶段、分层次的标准化推进策略，旨在促进光电协同网络技术的成熟与产业化应用。

第一阶段：物理层与链路层标准化

在这一阶段中，需要制定物理层的光交换设备接口标准，同时为了量化网络性能，还需建立光电性能基准与测试标准。

光交换设备的接口标准中，最重要的是控制接口规范，如重配置指令集、状态查询机制、性能参数定义等，以便清晰地以自动化方式控制重配置。另外，还需定义光链路重配置时间的参数标准，包括最大重配置时延、稳定性指标等参数。此外，为了保证跨层链路感知的良好实现，为动态重构提供基础支持，需定义链路层状态监测与反馈信息的格式。

光电性能基准与测试标准中，应包括吞吐量、延迟、抖动等关键指标的测量方法。同时，为了比较不同厂商产品，还要指定标准化的

测试用例与验证程序，并保证不同厂商产品之间的互操作性。

第二阶段：网络层与传输层标准化

在网络层，为了更快完成路由信息的收敛，可以采取双措并举的方法。在精简 BGP 的方面，基于 BGP 协议的精简版本，制定适用于光电协同网络的轻量级路由协议标准，重点优化收敛速度和资源开销，定义必要的属性集和探测机制；在高可用可扩展的 SDN 方面，制定 SDN 控制平面与光交换设备交互的接口标准，支持拓扑更新和路由推送功能，分层控制、负载均衡和边缘下沉的具体协议和接口等。同时制定路由协议性能的评估标准，包括收敛时间、控制开销、稳定性、近似最优度等方面。

在传输层，则分别实现前文中提到的光电协同高性能传输层协议，包括双态拥塞控制、动态多路径、拓扑有感的不公平流量调度等，并设计对应的向后兼容 API，支持应用层的调用。具体来说，在指定流量调度标准时，需要定义任务优先级评估（如先来先服务、最短完成时间优先等）、拓扑相似性和保证独占带宽的算法等；对于动态多路径，需要制定动态路径感知和流迁移决策的标准；而对于双态拥塞控制，则需借助光电链路切换的信令机制。同时结合上一阶段中的链路层反馈，实现传输层的拓扑有感设计。为了对传输协议后期进一步优化，还可以考虑建立传输层性能监控标准化框架，支持实时性能的分析。

第三阶段：应用层标准化

在这一阶段，需要为通用的智能通信库制定标准，如多种集合通

信算法的实现、动态选择机制和性能评估方法等，并利用链路层的感知接口，实现集合通信与物理链路的协同。为了配合传输层的不公平流量调度，需要定义资源竞争解决、分阶段执行和协同调度机制。

为了推动上层应用最大化利用光电协同网络及其新协议栈，需要指定应用适配指导标准。这包括分布式机器学习训推应用的网络优化指南，包括通信模式设计、参数配置、性能调优等；还包括不同智算应用的最佳实践标准，以及应用层网络感知的标准 API，为应用开发者提供协同参考。

在安全与可靠性方面，需要建立访问控制标准、网络可靠性评估标准，包括访问控制、数据加密、可用性指标、数据一致性等，以更好衡量其安全与可靠性。对于日常检查与灾难应对方面，需要指定管理接口标准、自动化运维标准、多厂商环境下的统一管理标准，以及灾难恢复的标准化流程，以指导如何在实践上改善安全与可靠性。

4.2 面向未来的研究与产业发展方向

未来，光电协同网络将在技术演进、应用拓展、产业生态等方面呈现出积极的发展趋势，为下一代智算基础设施奠定坚实基础。

技术演进方面，未来光电协同网络将超越单纯的传输功能，向光子计算与网络传输深度融合的方向演进。通过在光域直接进行矩阵运算、卷积操作等计算密集型任务，实现"计算即传输、传输即计算"的革命性架构。这种融合将显著降低数据在光电转换过程中的延迟和能耗，为大规模深度学习训练提供前所未有的性能优势。另外，下一代光电

协同网络将深度集成人工智能技术，实现从被动响应向主动预测、从人工配置向自主优化的根本性转变。基于大模型的网络智能体将能够理解业务意图、预测流量需求、自动生成网络配置，实现网络的自我管理、自我优化和自我修复。

应用拓展方面，依托光交换技术带来的大量端口，有望支撑更大规模的机器学习训练任务。对于超万卡规模的超大型 AI 模型训练，单次训练任务的数据交换量将达到 PB 级别。光电协同网络将提供 Tbps 级的聚合带宽和微秒级的通信延迟，使得万亿参数规模的大模型训练成为可能，推动 AGI（通用人工智能）技术的突破性进展。在智算场景以外，也可以为大规模超算任务提供支持：光电网络有望支撑飞机、汽车、船舶等复杂工程系统的数字化仿真设计。通过提供高带宽、低延迟的计算资源互联，使得工程师能够进行实时的协同设计和仿真验证，显著缩短产品开发周期。还可以连接分布在不同地理位置的大科学装置(如粒子加速器、射电望远镜阵列、基因测序中心等)，构建全球科研协作网络。光电协同网络将支持海量科学数据的实时传输和分布式处理，加速重大科学发现的产生。

产业生态方面，光电网络的发展将推动芯片与设备等上游产业升级，推动光子芯片与电子芯片的深度集成，形成新一代光电融合处理器。这类芯片将在单一封装内集成光子网络接口、电子计算单元、存储控制器等功能模块，为高性能计算提供一体化的解决方案。此外，网络调度算法和应用框架也将进一步改进，令开发者的学习成本进一步降低，实现更智能的网络资源分配和动态优化。

光电协同网络技术正站在历史性的发展机遇期，其未来发展将深刻改变智能计算的基础设施格局。通过技术创新、产业协作、标准化推进等多方面的努力，光电协同网络将从实验室走向产业化应用，从局部试点走向大规模部署，最终成为支撑人类社会数字化转型的核心技术基础。

面对激烈的国际技术竞争，我国应该抓住光电协同网络发展的战略窗口期，加大技术研发投入，完善产业发展政策，培育具有国际竞争力的企业和技术标准，在这一关键技术领域确立领先优势，为建设制造强国、网络强国、数字中国提供强有力的技术支撑。

参考文献

[1] 中国政府网.《国家数据基础设施建设指引》印发 将建全国数据“一本账”[EB/OL]. (2025-01-07) [2025-08-01]. https://www.gov.cn/lianbo/bumen/202501/content_6996610.htm

[2] 中国政府网.人工智能全球治理行动计划[EB/OL]. (2025-07-26) [2025-08-01].https://www.gov.cn/yaowen/liebiao/202507/content_7033929.htm

[3] 中国互联网络信息中心（CNNIC）. 第 56 次中国互联网络发展状况统计报告[R/OL]. (2025)[2025-07-21].

[4] NVIDIA.云与数据中心[EB/OL]. [2025-08-01]. https://www.nvidia.com/zh-tw/data-center/nvlink/

[5] 陆平静,董德尊等.高性能计算和数据中心融合网络研究综述[J].国

防科技大学学报,2023,45(04): 1-10.

[6] 任雄飞. 面向数据中心的光交换网络资源调度技术研究[D].北京:北京邮电大学,2024.

[7] 百度智能云. 百度云智一体白皮书[R]. 北京: 百度公司, 2023.

[8] 段晓东,李婕妤,程伟强,等. 面向智算中心的新型以太网需求与关键技术 [J]. 电信科学, 2024, 40 (06): 146-159.

[9] 武汉市互联网信息办公室. 人工智能为何如此耗电[EB/OL]. (2024-09-27)[2025-08-12].http://www.whwx.gov.cn/wlcb/wwtj/202409/t20240927_2460897.shtml

[10] Lumentum. Energy-Efficient AI Networks Lead to Dramatic Reduction in Environmental Impact. [EB/OL]. (2024-04)[2025-08-12]. https://resource.lumentum.com/s3fs-public/2024-04/eeainetwork-wp-oc-ae.pdf

[11] ZHANG Z, CHANG C, LIN H, 等, 2020. Is Network the Bottleneck of Distributed Training?[C/OL]//Proceedings of the Workshop on Network Meets AI & ML. Virtual Event USA: ACM: 8-13[2025-06-25]. https://dl.acm.org/doi/10.1145/3405671.3405810. DOI:10.1145/3405671.3405810.

[12] CHO M, FINKLER U, SERRANO M, 等, 2019. BlueConnect: Decomposing all-reduce for deep learning on heterogeneous network hierarchy[J/OL]. IBM Journal of Research and Development, 63 (6): 1:1-1:11[2025-06-25]. https://ieeexplore.ieee.org/document/886815

2. DOI:10.1147/JRD.2019.2947013.

[13] 曦智科技. 曦鲜事 | 国内首个光互连光交换 GPU 超节点——光跃 LightSphere X 发布[EB/OL]. (2025-07-28)[2025-08-12]. https://mp.weixin.qq.com/s/fKfoDMEgq3wSb6_UtTzDvQ