XXXX

星脉3.0：面向MoE模型训推一体的大规模高性能网络设计与实践

王亚晨[1]，夏寅贲[1]，王梓博[2]，曹培睿[2]，王智彬[2]
（1. 腾讯科技（深圳）有限公司，广东 深圳 518000；
2. 南京大学计算机软件新技术国家重点实验室，江苏 南京 210000）

摘　要：随着大模型架构向稀疏化混合专家模型（MoE）演进，训练及推理场景下的通信开销在端到端时延中的占比显著上升，通信性能逐渐成为制约系统性能的关键因素。针对大规模 MoE 训练与推理场景中 All-to-All 通信压力大、带宽及时延敏感以及运维复杂度激增等挑战，本文提出了一套软硬件协同的高性能网络基础设施解决方案。首先，在架构层面，本文利用光 Shuffle 技术构建扁平化的二级单轨网络，设计了星脉 3.0 网络架构，适配MoE All-to-All流量特征，显著提升了通信性能并降低了组网成本。其次，在通信软件层面，本文根据训练和推理中各个阶段的不同流量特点，分别进行针对性的 All-to-All 通信内核优化，利用以GPU为中心的下发技术及专家粒度的负载均衡技术，实现了适配训练与 Prefill 阶段的高带宽内核及适配 Decode 阶段的低时延内核，大幅降低了端到端时延。最后，在运维层面，本文利用AI Agent全面优化网络系统运维流程，实现了故障的主动预警与智能化交互诊断，保障了长周期训练的连续性与在线服务的高可用性。实验结果表明，该方案有效打破了 MoE 模型的通信墙，为万亿参数模型的大规模训练与在线服务提供了统一的高性能、高可靠系统底座。

关键词：混合专家模型；星脉3.0网络架构；All-to-All 通信；智能运维；训推一体
中图分类号：TP393
文献标志码：A
doi: 10.11959/j.issn.1000−0801.

Astral 3.0: Design and Practice of High-Performance Network Infrastructure for Large-Scale MoE Training and Inference

WANG Yachen[1], XIA Yinben[1], WANG Zibo[2], CAO Peirui[2], WANG Zhibin[2]
1. Tencent Technology (Shenzhen) Co., Ltd., Shenzhen 518000, China
2. State Key Laboratory of Novel Software Technology, Nanjing University, Nanjing 210000, China

Abstract: With the evolution of large model architectures towards the sparse Mixture-of-Experts (MoE), the proportion of communication overhead in end-to-end latency was observed to rise significantly in both training and inference scenarios, and communication performance gradually became a critical factor constraining system performance. To address challenges such as heavy All-to-All communication pressure, sensitivity to bandwidth and latency, and

收稿日期：XXXX−XX−XX；修回日期：XXXX−XX−XX

surging operational complexity in large-scale MoE training and inference scenarios, a high-performance network infrastructure solution based on hardware-software co-design was proposed in this paper. First, at the architecture level, the Astral 3.0 network architecture was designed by utilizing Optical Shuffle technology to construct a flattened two-layer single-rail network. This architecture was adapted to the All-to-All traffic characteristics of MoE, significantly improving communication performance and reducing networking costs. Second, at the communication software level, targeted All-to-All communication kernel optimizations were performed based on the distinct traffic characteristics of various stages in training and inference. By utilizing GPU-centric task dispatch technology and expert-granularity load balancing technology, high-bandwidth kernels adapted for training and Prefill stages, as well as low-latency kernels adapted for the Decode stage, were implemented, which drastically reduced end-to-end latency. Finally, at the operations level, network system operational workflows were comprehensively optimized using AI Agents, and proactive fault warning along with intelligent interactive diagnosis were achieved, ensuring the continuity of long-term training and the high availability of online services. Experimental results demonstrated that the communication wall in MoE models was effectively broken by this solution, providing a unified, high-performance, and highly reliable system foundation for the large-scale training and online service of trillion-parameter models.

Key words: Mixture of Experts (MoE), Astral 3.0 Network Architecture, All-to-All Communication, AIOps, Inference System

0 引言

当前，人工智能技术正加速向通用人工智能（AGI）演进，模型参数规模呈现指数级增长态势。为了在突破算力与能耗墙的同时提升模型性能，混合专家模型（Mixture of Experts, MoE）凭借其稀疏激活的特性，已逐渐取代传统的稠密 Transformer 架构，成为超大规模模型设计的主流范式。然而，MoE 架构在显著降低单次推理计算量的同时，也引入了复杂的专家动态路由机制。这种机制导致节点间的数据交换量及数据交换频率大幅增加，使得通信开销在推理总耗时中的占比显著提升，通信效率已成为决定系统整体性能的关键要素。

特别是在大规模集群的训推一体场景中，MoE 模型特有的 All-to-All 通信模式带来了双重挑战。一方面，在训练阶段，超大规模的 All-to-All 集合通信在瞬时带来巨大的吞吐压力，极易在传统的多层级网络架构（如 CLOS）中引发哈希冲突与网络拥塞，严重制约了分布式训练的线性加速比；另一方面，推理阶段（尤其是 Decode 阶段）主要由高频的小数据包构成，对时延极为敏感。然而现有的工业标准通信库（如 NCCL）难以有效应对此类突发流量，导致昂贵的 GPU 算力因等待网络数据而处于空闲状态。此外，随着集群规模迈向十万卡级别，网络设备的维护复杂度激增，传统依赖人工规则的被动运维模式已难以保障系统的高可用性。

针对上述挑战，本文提出了一套面向大规模 MoE 训推一体集群的软硬件协同高性能网络基础设施解决方案，从物理网络架构、底层通信库优化以及智能运维体系三个维度展开论述，旨在构建高性能、低时延且高可靠的基础设施底座。本文的主要贡献如下：

● 首先，在物理架构层面，本文设计了星脉 3.0 网络架构。针对传统多层网络扩展性差与跨轨通信效率低的问题，利用光 Shuffle 技术构建了扁平化的二级单轨网络。该架构成功将十万卡集群整合为统一的逻辑通信域，显著减少了数据转发跳数，提高了通信性能，降低了组网成本。

● 其次，在通信软件层面，本文设计了面向

训推场景自适应的 All-to-All 通信内核优化方案。针对训练与推理 Prefill 阶段的大包高吞吐特征，以及推理 Decode 阶段的小包低时延特征，通过以GPU为中心的高效任务下发技术以及专家粒度的精细化负载均衡技术，解决了带宽利用率不足与网卡负载不均的问题。实验表明，该方案在提升带宽利用率的同时，大幅降低了端到端通信时延。

● 最后，在运维层面，本文引入了 AI Agent 以全面优化网络系统运维流程。通过构建智能化诊断分析体系，实现了从被动响应到主动风险预警的转变，有效解决了大规模网络环境下故障定位难、效率低的问题，为集群的长期稳定运行提供了智能化保障。

1　大模型技术的演进与通信挑战

1.1　稀疏化MoE架构成为大模型主流范式

根据近期工业界披露的数据显示，大模型推理服务的负载正处于爆发式增长的前夜。据国家数据局统计，我国日均 Token 消耗量在过去一年半里增长了 300 多倍 [1]；头部科技企业如谷歌，其 Token 消耗量也在一年内实现了百倍级增长 [2]。面对指数级上升的吞吐压力，模型架构的稀疏化（Sparsification）已成为打破算力供需矛盾的核心演进趋势。这一趋势在 Transformer 的两大核心组件中均表现显著：一方面，注意力机制（Attention）正通过滑动窗口等稀疏技术降低长序列计算开销 [3][4]；另一方面，前馈神经网络（FFN）也正迅速向稀疏的混合专家（MoE）架构转型。这种全方位的稀疏化演进，打破了传统稠密（Dense）模型中计算复杂度与参数规模的线性约束，使得在资源受限下部署超大规模模型成为可能。

在这一稀疏化浪潮中，混合专家模型（MoE）凭借其成熟的稀疏激活机制，已确立为超大规模模型设计的主流范式。如图 1-1 所示，该架构通过路由模块（Router）将输入 Token 动态路由至少数几个 (top_k) 专家（Experts），实现了 FFN 层面的计算量与参数规模解耦。Fedus 等人提出的 Switch Transformer [5] 证实了 MoE 模型容量的线性扩展能力，谷歌 GLaM [6] 验证了仅需激活约 8% 参数即可实现低能耗推理。当前，DeepSeek [7]、Qwen [8]、Mistral [9] 等主流开源大模型均采用了 MoE 架构。

1.2　MoE模型训练和推理中的通信瓶颈

在大规模训练与推理部署中，专家并行

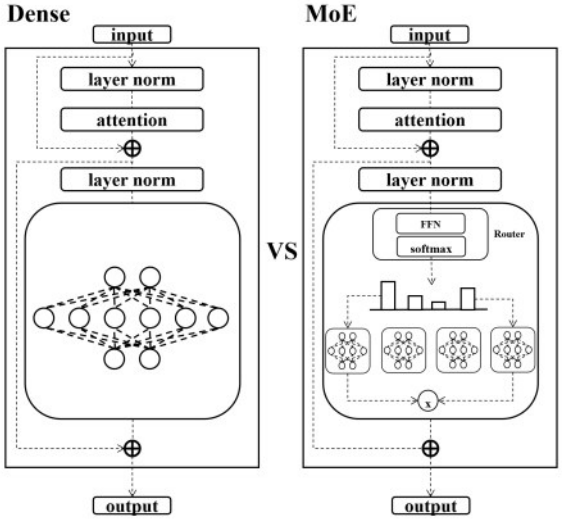

图 1-1　稠密与MoE架构对比图

（Expert Parallelism）通常是容纳海量模型参数的必要手段。在该并行模式下，专家参数被分片存储于不同 GPU 中，Token 对专家的动态选择机制引发了设备间不规则且全连接的数据交换，即 All-to-All 通信。如图 1-1 所示，该通信的时延直接暴露在推理的关键路径上，且难以被计算过程有效掩盖（Overlap）。更为严峻的是，All-to-All 通信的开销极为显著。BigMac [10] 的系统化评测表明，在 32 GPU 的专家并行规模下，随着 top-k 从 1 增加至 8，All-to-All 通信在端到端推理延迟中的占比由 51.2% 飙升至 90.6%。这不仅严重拖慢了推理响应速度，更成为了限制分布式训练集群线性扩展比（Linearity）的核心瓶颈。

这一通信瓶颈的根源，源于 MoE 架构在训练与推理阶段对网络提出的双重挑战：既要满足训练时的超大规模吞吐，又要保障推理时的极致低时延。在训练阶段（及推理 Prefill 阶段）系统通常采用大批次（Large Batch）模式，这导致超大规模的瞬时 All-to-All 集合通信流量。在传统的 CLOS 多层级网络架构中，这种激增的突发流量极易引发哈希冲突与网络拥塞，导致网络二分带宽（Bisection Bandwidth）利用率下降，严重制约了万卡集群的训练效率与扩展性。在推理的 Decode 阶段系统每步仅处理少量 Token，导致通信包大小相较于训练阶段下降了两个数量级。这种高频小包的流量特征，使得面向大消息吞吐优化的工业标准 NCCL 通信库难以发挥效能，其固有的内核调度开销成为了显著阻碍；与此同时，底层的 InfiniBand 和 RoCE 等协议栈也难以高效适配此类突发流量，导致昂贵的 GPU 算力因等待网络数据而出现空闲时间，利用率降低。

针对 MoE 架构固有的通信挑战，现有研究已从模型结构设计与系统调度实现两个层级展开了广泛探索，并取得了显著的量化收益。在模型结构与算法层面，核心思路是通过降低通信数据维度或减少跨设备路由频次来缓解带宽压力。BigMac [10]通过在专家层引入“降维-通信-升维”机制（DCCA），使 All-to-All 在低维空间完成，实测在推理吞吐上实现了最高约 3.11 倍的提升。MoE++ [11]引入零计算专家（Zero-Computation Experts）以实现计算的“本地化”，在同等规模下将专家前向吞吐提升了 1.1 至 2.1 倍。这些工作证明了通过结构性剪枝与压缩，可以显著降低通信的数据量。在系统实现与并行策略层面，研究重点在于优化通信拓扑与隐藏延迟。微软推出的 Tutel [12]提出了 2D 分层 All-to-All 算法，通过感知物理拓扑进行自适应并行策略切换，在大规模集群（2048 GPUs）下实现了相对于基线约 5.75 倍的加速。

然而，尽管上述算法与系统层面的优化有效降低了通信开销，但物理网络层面的硬性约束依然存在。结构性压缩虽然减少了数据传输总量，却导致通信粒度进一步碎片化，使得网络流量特征显著向高频小包演变；而系统层面的分层调度虽能提升吞吐，却无法从物理上消除跨轨通信带来的高跳数与拥塞风险。特别是在训推一体的超大规模集群中，单纯依赖上层软件调优已难以突破由传统多层级交换架构及通用协议栈所构筑的物理性能天花板。

综上所述，稀疏化 MoE 已成为突破稠密模型扩展瓶颈的核心技术路线，但其训练扩展性与推理实时性均受到通信开销的严厉制约。随着大规模分布式集群的普及，基础设施的优化重心正从单一的算力资源扩容，转向对计算-通信比（Comm-to-Comp Ratio）的系统级调优。针对当前 MoE 存在的通信屏障，本文将分别从星脉 3.0 网络架构的物理组网优化、通信库的性能优化、以及基于智能运维的系统保障体系三个维度，系统阐述面向大规模 MoE 训推一体的高性能网络基础设施设计与实践。

2　星脉3.0网络架构

在星脉2.0[13]及其之前的网络架构中，其架构针对的目标主要是稠密模型（Dense Model）。随着MoE模型的出现，流量模式与先前的模型相比发生了一定的变化，表现为引入了更多的All-to-All通信，这在星脉2.0上会引入大量的跨轨通信，使得流量大量通过核心层带来路径变长与拥塞。为了适配MoE模型的流量模式，我们提出了星脉3.0架构。其采用单轨架构，引入光交换设备将交换机物理端口拆分成4个通道，并通过自研硬件底座与协议实现在性能、成本和效率上的显著提升。

2.1　面向MOE通信特征的High Radix单轨道扁平网络架构

大规模语言模型正经历从稠密模型（Dense Model）向稀疏模型（比如混合专家模型MoE）的转变，已经进入指数性扩张的阶段，模型参数从千亿迈向万亿，数据中心集群规模也从万卡级向十万卡级推进。MoE模型中核心的专家并行机制引入了全局性的All-to-All通信，这种模式下的输入token会动态激活选择专家，导致数据需要在节点之间进行无规则、全连接的数据交换，这对数据中心网络的可扩展性、带宽、延迟和任意两点间的通信效率都构成了挑战。

过往多轨道架构[14][15][16]将集群划分成多个通信域，跨域通信会引入上层网络，增加通信跳数，从而拖慢通信性能。在我们的测试中，流量跨轨通信会额外引入10us的通信时延。这在推理Decode阶段的多次通信中，会累积产生1.2ms的额外通信时延。在AI agent等低推理时延场景（词间时间10ms SLA），占比超过10%。在万卡级别的分布式训练中，多轨架构的跨轨通信同样限制了All-to-All的集合通信效率，降低了训练任务的线性加速比。

在此背景下，我们提出星脉3.0网络架构。该架构旨在利用下一代102.4Tbps高性能交换芯片的能力构建扁平化网络。尽管该芯片在逻辑上已具备支持512个200G端口（High-Radix）的能力，但在物理设备的工程实现上却面临严峻挑战。具体而言，若通过增加物理端口数量来实现High-Radix（例如将主流的64x800G形态调整为256x200G），单机物理高度将从2U激增至6U。这一物理尺寸的膨胀显著拉长了交换芯片到出端口的走线距离，使得链路更加复杂，带来了极大的信号完整性（SI）挑战与整机硬件设计复杂度。此外，在能效与成本维度，单个800G光模块的功耗比4个200G光模块低30%，且64x800G形态的单机成本远低于256x200G形态。

为了解决芯片逻辑高Radix能力与物理设备实现（散热、体积、SI、功耗及成本）之间的矛盾，星脉3.0引入了光Shuffle技术作为关键解法。我们保持了交换机在64x800G这一高效形态下的物理设计，通过结合外部的光Shuffle互联与交换芯片的内部通道拆分能力，成功将AI集群网络从典型的三层CLOS架构压缩到更为扁平的两层架构，从而构建了一个覆盖10万GPU的、逻辑上统一的单轨通信域。这种两级架构不仅显著减少了网络设备数量、简化了网络管理运营、降低了TCO，更重要的是，通过减少数据转发跳数和优化数据通信路径，有效降低了网络时延性能，并提升了网络吞吐量和整体稳定性。

2.2　架构设计与组件剖析

星脉3.0架构核心设计：利用光shuffle等物理层交换设备，让交换机连接更多的对端设备，从而提高组网规模。

2.2.1　利用光交换互联扩大组网规模

在传统连接中，一个高速网络端口作为一个不可分割的整体，只能固定地连接到一个对等端口上，大大限制了灵活性。星脉3.0的核心洞察在于将每个交换机物理端口拆分成多个独立的通道（Lane），这样可以扩大组网规模。这种设计

显著提升了交换机的端口密度（Radix）：在物理端口带宽恒定的前提下，通过增加 Lane 的数量，单台汇聚层交换机（Leaf）能够连接的接入层交换机（ToR）数量成倍增加，从而在接入层下挂服务器数量不变的情况下，几何级数地扩大了集群的整体组网规模。

在上述原理的支撑下，星脉3.0将每个800G物理端口拆分成4个独立的200G通道。在实现层面，该架构利用光交换（Shuffle）技术对信号路径进行物理层重组，使得源端口拆分出的 4 个 Lane 能够被精准路由至不同的目的端设备。如图 2-1 所示，这种高带宽下的细粒度互联设计，有效支撑了大规模集群的扁平化组网需求。

在工程落地与可靠性设计方面，虽然引入光 Shuffle 设备（Shuffle Box）看似增加了物理层的复杂度，但实际上我们将复杂度从有源设备转移到了无源或低故障率的物理层设备上。星脉 3.0 中的光 Shuffle Box 内部主要由高精度的无源光纤交叉矩阵构成，无电子元器件，其平均无故障时间（MTBF）远高于传统有源交换机。针对可能的单点物理故障（如光纤损坏），架构在设计上天然具备故障隔离能力：一个 Shuffle Box 的故障仅影响其连接的特定 4 个通道（Lane），而非整个网络节点。结合上层的多路径路由算法，流量可毫秒级切换至其他健康的物理通道，确保了系统的整体高可用性。此外，在大规模布线管理上，我们采用了预连接光缆（MPO）与高密度配线架方案，将 Shuffle 互联封装在标准机柜内，大幅降低了现场施工的复杂度与人为故障风险。

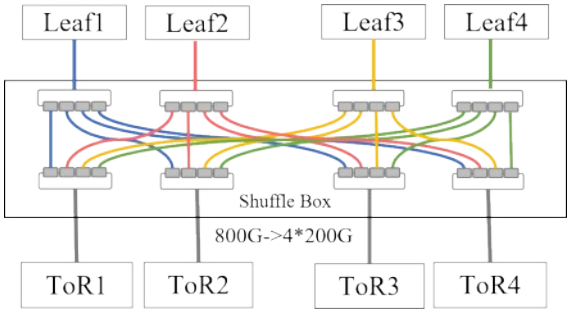

图2-1　星脉3.0 光交换（shuffle）组件

2.2.2　从多轨到单轨的技术演进

如图2-2所示，星脉2.0采用多轨网络架构，支持ToR（Top of Rack）交换机的同轨聚合，即为了在同一个轨道上实现最大数量的GPU，将每个块（Block）中的两个同轨ToR交换机分别连接到第二层的两组64个聚合交换机（Aggregation Switch），最终使得一个Pod可以支持64K GPU同轨通信。然而，MoE模型中特征性的全局All-to-All操作不可避免地会引入大量跨轨通信需求。尽管扩大单轨GPU规模能在一定程度上提升通信局部性，但无法从根本上消除跨轨流量。此类流量被迫经过核心层（Core Layer）转发，导致数据传输路径显著延长及网络时延增加，成为制约大规模集群性能的关键瓶颈。

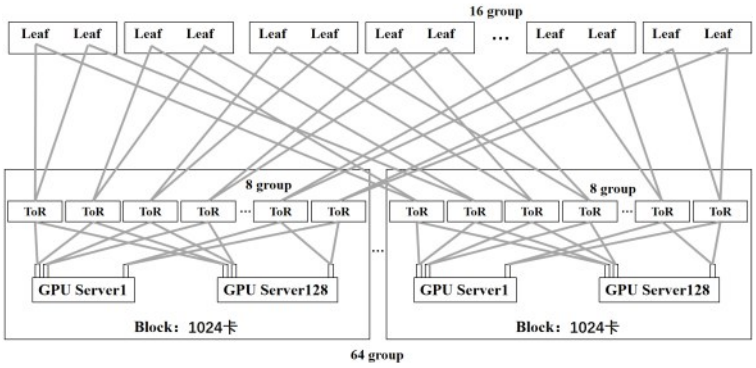

图2-2　星脉2.0架构图

在MoE架构模型逐渐占据主导的背景下，我们针对其特有的通信模式对网络架构进行了系统性的重构。为了清晰阐述星脉 3.0 的设计原理与拓扑结构，本节将首先解析构成该架构的三大核心组件：GPU 服务块（Block）、光交换模块以及接入聚合层。

● **GPU服务块（Block）**作为集群的基本模块，一个Block 包含 256 卡，如图中由 32 台 8-GPU 服务器构成，这些服务器还需接入 ToR（Top of Rack）。这相较于星脉 2.0（1024 卡/Block）更细粒度的设计，赋予了系统更好的故障隔离能力与更平滑的线性扩展能力。每个GPU通过专属高速网络端口接入ToR，确保在源头上有充足的网络带宽。

● **光交换模块**用于实现物理端口拆分成多个通道这一核心技术洞察。与原先星脉 2.0 中 ToR 直连到Leaf switch不同，星脉3.0采用了光交换，一个交换盒（Shuffle Box）中采用 fullmesh 全互联，盒中下层到上层全互联，下层节点之间不互联。在不采用光shuffle前，Leaf switch一个800G的端口只能连接到一个800G的ToR端口。在用光shuffle之后，Leaf switch一个800G端口实则为聚合带宽，它由4个200G的链路带宽组成，链接到光 shuffle 交换机后可以分别连到不同的 4 个ToR。对高速端口进行带宽细粒度的切分可以在底层硬件不变的前提下仅引入shuffle就可以将整个二层网络的终端接入能力提升4倍。

● **接入聚合层**即传统ToR-Leaf-Spine结构中的Leaf层设备，它汇聚了来自ToR的数据。

如图2-3所示，星脉3.0架构采用单轨架构，GPU服务器内的每个GPU网卡都会接入同一个ToR，ToR经过光交换模块连接至上层的Leaf switch，其中光交换模块可以将一个800G交换机端口拆分为4个200G通道（Lane）以提供更灵活的路径选择。在这种大二层网络架构下，跨轨的概念及其带来的性能开销就被消除了，这种扁平化的网络也解决了东西流量的问题，为MoE模型大量的All-to-All操作提供了优越的网络基础。

2.3　架构优势以及性能、成本与效率收益

扁平的二级架构从根本上减少了数据包转发跳数，结合光Shuffle对通信路径的优化，为MoE模型的专家并行（EP）提供了有效的网络支撑，既满足了MoE训练对大规模吞吐的需求，也为在线推理提供了极致的低时延保障。大二层的组网减少了网络路径长度，有利于提升MoE通信性能。经过实验可得，All-to-All通信时延降低了20%。**在性能收益方面**，为了量化评估星脉 3.0 架构在低时延通信上的优势，我们在两台同 Block 的 GPU 服务器（共 16 张 H20 加速卡）上进行了 DeepEP Low Latency Benchmark 微基准测

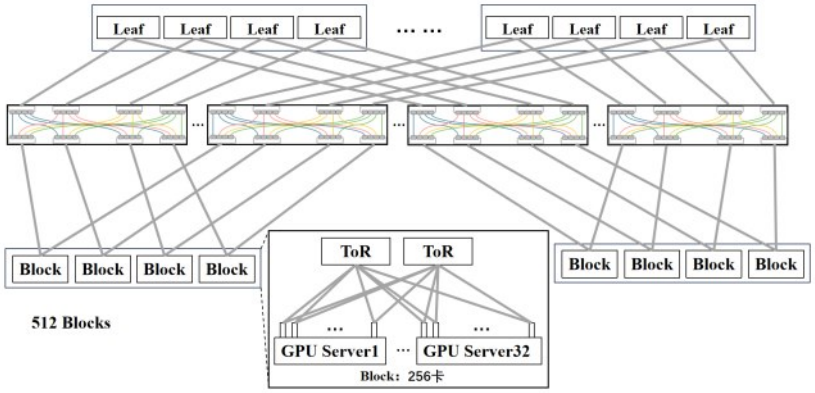

图2-3　星脉3.0架构图

试。实验设置 All-to-All 通信的 Batch Size 为 32，对比了星脉 2.0（多轨架构）与星脉 3.0（单轨扁平架构）的通信完成时间。测试结果显示，**Dispatch 阶段的通信时延从 56us（星脉 2.0）降低至 48us（星脉 3.0），Combine 阶段的通信时延从 77us 降低至 65us**。这一约为 15% 的基础时延优化，主要得益于单轨架构下更短的物理链路与更高效的转发路径，有效减少了小包通信在交换设备内部的排队与仲裁开销。此外，单轨架构减轻了集群内部GPU之间的通信壁垒，使得集群性能可以随着GPU数量的增加而实现近乎线性的扩展，为未来的百万卡集群铺平了道路。

在硬件底座上，采用自研TH6 102.4T芯片；高速互联采用自研800G BR8和800G DR4.2+。800G BR8可以覆盖楼内互联距离；800G DR4.2+实现了楼间1km互联，相比商用800G 2*FR4方案，成本降低30%。自研算力网卡ASIC化，实现了Scale out 800G接入。互联速率、交换芯片容量均提升了一倍，持续优化成本和功耗。

在组网拓扑上，接入带宽由400G变为800G，各层级的规模均增加了一倍。在成本效益方面，基于128K GPU集群的详细物料清单（BOM）核算，相比星脉2.0方案，星脉3.0通过单轨高Radix组网大幅减少了汇聚层交换机与光模块的数量。具体而言，交换机及光模块等核心物料数量减少了69%，整体网络系统的采购成本（TCO）降低了48%。这不仅大幅降低了建设成本，也显著降低了后续的机房功耗与维护压力。在交付与成本上，以128K集群为例计算，相比星脉2.0，物料数量减少69%，采购成本减少48%。它采用了弹性设计，可以满足不同规模的各种机型，比如NV B20、GB200、910C、紫霄超节点等弹性异构接入。

如图2-5所示，星脉3.0相比星脉2.0，在接入带宽、Block规模、二层Cluster规模、三层Pod规模、三层Cluster规模均有一倍的增加，可以明显的发现架构升级、物理端口进行多通道拆分、硬件与协议的综合支持为星脉3.0提供了有力保障。

综上所述，星脉3.0直面十万卡集群的网络互连挑战，通过协同利用102.4Tbps/512-port Radix交换芯片和光Shuffle的互联，成功构建了一个两级扁平、逻辑单轨的超大规模网络。它不仅消除了跨轨通信瓶颈，还实现了性能提升和成本节约。

3　以GPU为中心的高性能通信库

针对MoE架构特殊的All-to-All通信流量特征，我们为不同的任务阶段（训练/Prefill或Decode）分别设计优化了高带宽和低延迟的通信内核。我们不再采用传统的CPU主导的通信模式，转而利用GPU的高并发特性来提高网卡的利用率，并针对多端口网卡做了针对性的优化，同时从网络层面降低了专家负载不均对通信性能的负面影响。

3.1　All-to-All通信模式

MoE架构的核心在于专家并行机制引入的All-to-All通信。在大规模训练与推理的全生命周

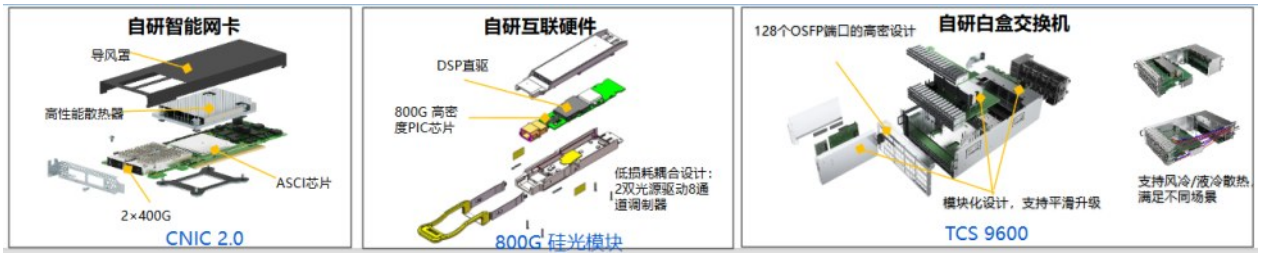

图2-4　自研智能网卡、自研互联硬件、自研白盒交换机结构图

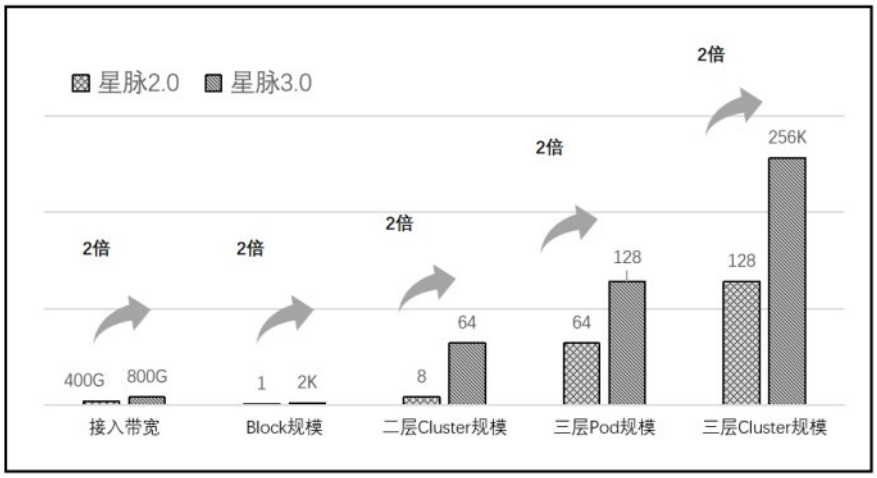

图 2-5　星脉 2.0 和星脉 3.0 接入和集群规模对比

期中，通信模式呈现出明显的阶段性特征。

对于推理任务，Prefill 阶段一次性读入用户请求输入的全部上下文 Token，生成输入文本的所有 KV Cache 并计算第一个 Token；Decode 阶段则是逐 Token 自回归生成，每次仅处理一个输出结果。值得注意的是，MoE 训练阶段的计算与通信特征与推理的 Prefill 阶段高度相似，均表现为大批次、高吞吐的数据交互模式；而 Decode 阶段则表现为高频、小包的低时延模式 。

无论是在训练、Prefill 还是 Decode 阶段，All-to-All 通信都包含 Dispatch 和 Combine 两个关键步骤。如图 3-1 所示，Dispatch 过程是将各个 GPU 上负责的 Token 分发给通过 MoE 路由算法选定的部分专家，各个专家执行 Expert FFN 操作；Combine 过程则是将各个专家计算的结果返回给发送方进行加权聚合以得到最终结果。

3.2　高带宽通信内核优化

对于训练及推理的 prefill 阶段，所有 prompt 的 tokens 需要一次性完成全量前向计算，这会在多 GPU 的专家并行场景中触发大规模的 all-to-all 数据交换，因此这个阶段对于通信库的核心需求是能够提供足够大的带宽，从而避免通信成为吞吐瓶颈，整体架构如图 3-2。然而，在星脉网络架构的实际生产环境中，存在两点问题：第一，如果使用传统的 CPU 主导的 RDMA 通信，会由于计算单元缺乏足够的并行度，出现网卡带宽利用率不满载的问题；第二，在星脉网络架构中，RDMA 网卡是多端口架构，如果不加以规划会出现流量负载不均问题，严重降低网卡有效利用率。

在介绍具体技术之前，首先给出一些术语定义：

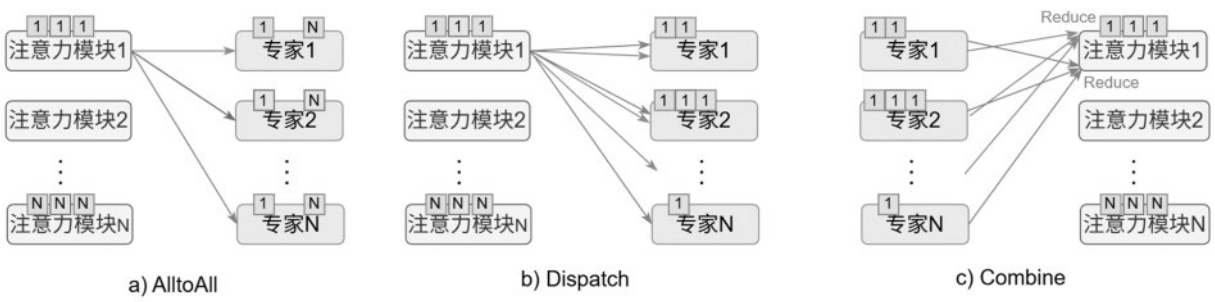

图 3-1　AlltoAll, Dispatch 和 Combine 通信模式差异

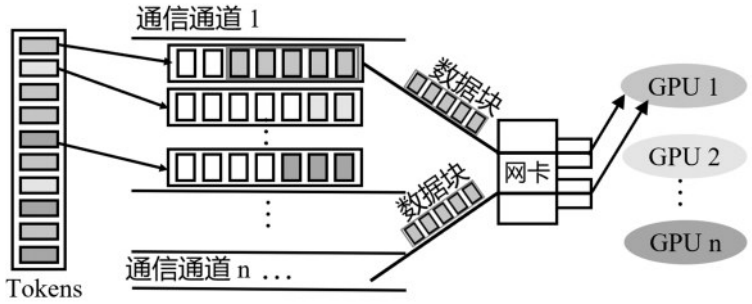

图3-2　高带宽通信内核整体架构

● 通信通道：为了对下层物理资源做逻辑抽象，提出通信通道的概念。每个通道被分配定额的GPU计算资源SM（streaming multiprocessor），每个通道内部数据流串行处理，通道间并行处理。每个通道内的token被聚合成数据块以最大限度提升网卡带宽利用率。每个通信通道内要发送的数据总量近似相等。

● QP（queue pair）：RDMA技术中，QP是由发送队列和接受队列组成的通信端点，用于RNIC上建立可靠、有序或无序的数据传输路径。在本设计中，每个QP被安排负责等量的通信通道的传输。

● WQE（work queue element)：RDMA技术中，WQE是提交到QP上的操作描述符，用于指示RDMA网卡执行特定的数据传输或内存访问任务。

3.2.1　基于高并发的任务下发

如果采用CPU主导的IBRC（Infiniband Reliable Connection）通信模式，数据传输指令会首先被打包成DMA（Direct Memory Access）请求并被放入一个DMA队列，一个CPU代理线程通过DMA机制取出一个请求并生成一个WQE，接着被串行写入一个QP，这种串行机制严重制约了网卡带宽的高效利用。

具体来说，如图3-3，对于网卡硬件而言，处理每个WQE需要一段硬件的固有延迟T_{HWI}和数据的实际传输时延T_{data}，如果使用CPU代理线程，串行的下发WQE会导致T_{WQE}（两个WQE的下发间隔时间）过长，$T_{WQE} > T_{HWI} + T_{data}$，进而导致硬件处理时间存在间隙$T_{inter}$，这就会最终使得网卡带宽无法打满。

就上述问题，本文提出的解决方案包含两个优化点。首先，为了减少T_{inter}，本文将IBRC替换成IBGDA（Infiniband GPU Direct Async），即GPU主导的通信，利用GPU的大规模并行能力来批量下发WQE，如此，每个通信通道可以使用自己的GPU资源SM同时的下发WQE，不会再出现CPU代理线程的串行处理问题。其次，即使T_{inter}被完全消除，由于T_{HWI}的存在，网卡的带宽仍然不能得到完全有效的利用。对此，我们在每对通信的GPU实体间建立多个QP连接。当一个网卡上建立多个QP的时候，网卡会采取一种Round-robin的策略：对QP进行轮流轮询，如果当前QP里存在需要发送的数据，那么网卡会为

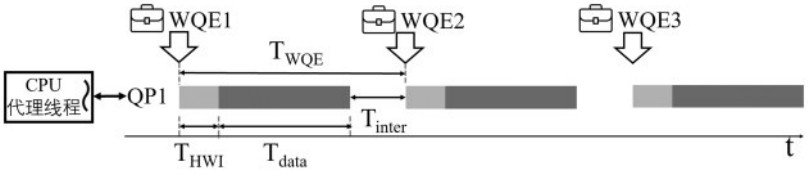

图3-3　IBRC通信模式的单线程任务下发过程

这个QP分配一个固定时间片长度的时间，让它传输数据，然后切换到下一个QP。如图3-4，这种方式就可以使用其他QP的有效数据传输时间t_e来掩盖住T_{HWI}，进而实现网卡带宽的满效率利用。

关于计算资源的占用分析： 理论上，通信库在调用 RDMA 传输前需执行量化与数据准备等预处理。在 IBGDA 模式下，负责预处理的线程束（Warp）直接构建并提交 WQE；而在 IBRC 模式下，同一线程束需通知 CPU，再由 CPU 构建 WQE。这表明 IBGDA 并不比 IBRC 占用更多的 GPU 线程束，且能更快释放 CPU 资源。实验表明，在 DeepEP 的 HB 内核中，采用 IBRC 模式时约占用 24 个 SM，切换至 IBGDA 模式后数量基本持平，此开销主要源于缓冲区数据迁移与排序，而非通信模式本身。经过我们进一步的代码优化，目前仅需 8 – 12 个 SM 即可使带宽达到饱和状态。相对于现代 GPU（如 H800 拥有 100+ SM）的总资源而言，这种占用对模型核心计算任务的影响微乎其微。

我们对方案进行了测试，测试环境中，每个GPU使用一个单端口的400Gbps的网卡，在不同EP数（即有专家分布的GPU数量）下，我们的方案都取得了带宽提升，如图3-5，在不同情况下，带宽提升在8.7%到31.8%之间。

3.2.2　端口号预规划

虽然3.2.1节中，高并发的任务下发已经可以让单端口网卡大幅度提升带宽利用率，但是现在大部分现代大规模数据中心中使用的都是多端口网卡（包括我们的星脉网络架构），这就带来了一个额外的挑战：如果没有精确地编排规划，那么流量会在网卡的多个端口上出现严重的负载不均问题，进而导致网卡带宽无法得到有效利用。

为了实现多端口网卡的负载均衡，我们在上

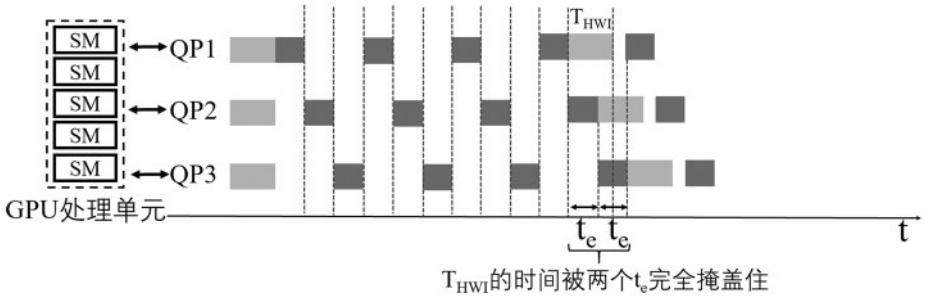

图3-4　IBGDA通信模式的多线程任务下发过程

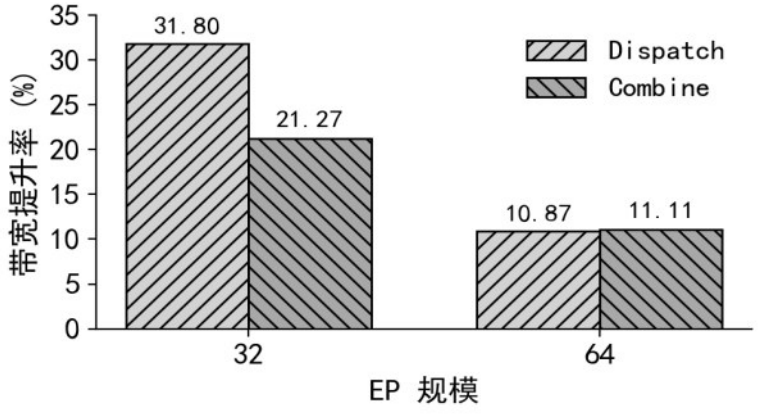

图3-5　高带宽内核高并发优化在不同EP规模下Dispatch和Combine的带宽提升率

文所述的一对通信实体建立多QP的基础上，做了端口号预规划（SPA，Source Port Pre-Allocation）的设计。一张网卡配备多个物理端口，网络流量基于五元组哈希选择一个端口进行发送，五元组的结构为<源ip地址，目的ip地址，源端口号，目的端口号，L4协议类型>，在一对通信实体之间的流量，源ip和目的ip总是相同的，目的端口号在RDMA标准中也被固定为4791，L4协议也是确定的，所以可以修改的只有源端口号。我们预先计算出多个特定的端口号，以保证五元组哈希值会被均匀的分布在不同的物理端口上。由于我们给每个QP分配的通信通道数是均匀的，每个通信通道内要发送的数据量是相同的，所以每个QP要发送的数据量近似相同。我们将QP均匀的映射到这4个特定的端口号上，进而保证了发送的数据总量在网卡的多个物理端口上是均匀的。

我们在配备了双端口网卡的集群上测试了端口号预规划的实验效果，没有端口号预规划时，带宽仅为有端口号预规划的53.3%，这是因为如果没有端口号预规划，数据在双端口网卡上几乎只走一个端口，浪费了近一半的网卡可用带宽。

3.3 低延迟通信内核优化

低延迟通信内核是专门为了推理中decode部分设计的，如图3-7。推理的decode阶段，用户体验的评判指标为TPOT（Time Per Output Token），所以降低每个token传输过程的时延是这个阶段的优化重点。由于当前LLM的自回归特性，token串行生成，即一次只生成一个token，所以同一时刻每个请求的生成过程中AlltoAll通信都只需要完成一个token的传输任务，所以传输带宽并非主要限制因素。在这种低延迟的要求下，如果仍然采用高带宽通信内核传输的方式，在发起传输之前将多个token打包成一个大数据块再发送，这个打包的过程会增加端到端时延，故不再打包，每个token各自传输。同时，低延迟内核的数据传输粒度是每个token直接以各个专家为目的地，这就把通信实体的粒度缩小了，进而需要给每个专家（通信端点）建立QP。

尽管取消打包操作会导致报文数量（PPS）激增，但现代高性能网卡（如ConnectX-7）具备亿级PPS的处理能力，且IBGDA模式大幅降低了主机侧CPU的中断与处理开销。我们的压力测试显示，在典型MoE推理负载下，当前的PPS峰值仍远低于网卡处理极限，未出现缓冲区溢出或丢包现象，因此暂无需引入额外的复杂流控机制。

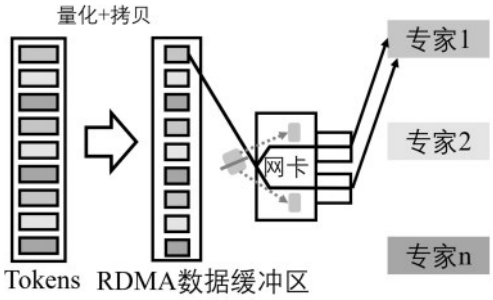

图3-7 低延迟通信内核整体架构

在decode阶段，通信库的设计有两大挑战：第一，相较之高带宽通信内核，低延迟通信内核的token是直接被发送给分布在各个GPU上的专家，而非统一发送给某个GPU再进行内部转发，这会导致给各个专家不均衡的数据量分布（如图3-8）。第二，给每个专家分配一个QP的做法，已经让在高带宽内核中存在的并发度不足的问题天然解决了，现在的瓶颈落在了网卡硬件上，即硬件处理每个WQE的时间过长，让整体端到端时延无法达到最优。为了解决这些挑战，我们提出了两点优化，在保证并发度的同时最大程度降低端到端的时延。

3.3.1 专家粒度的端口均衡

在decoding阶段，每个token是直接被发送给分布在GPU上的各个专家的，没有聚合步骤。而哪个token发送哪个专家的对应关系，是由

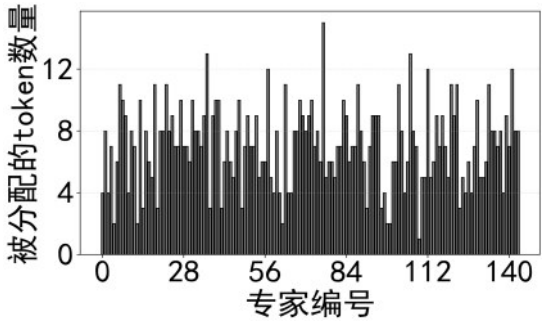

图3-8　Decode阶段专家数据量分布不均

MoE模型的gating function决定的，这很可能是不均匀的。即使使用了端口号预规划技术，也无法保证发送给每个专家的数据量均衡，进而累计起来在网卡的多个端口流量也并不均衡。

为了解决这个问题，我们进一步切分了发送给每个专家的token。以双端口网卡为例，我们为每个专家通信端点建立2个QP，同时根据端口号预规划原则，有一个从QP到网卡端口的映射。每个token被等分成两块，每块通过一个固定的QP发送，进而可以保证给每个专家通信端点的两个QP数据量是完全均衡的，再结合上述端口号预规划技术，不难得到最终结果是在两个网卡端口上实现了完美的均衡。

我们对此进行了实验验证。如图3-9展示了网卡两个端口利用率的比例，图中的端口利用率比例指的是取双端口网卡两个端口的有效数据带宽比值，比值越小表示两个端口利用程度越均衡。可以看到，在没有采用基于专家的负载均衡技术之前，两个端口利用率始终不均衡，最大可达到6.21的比值。而在使用基于专家的负载均衡技术之后，两个端口的使用率始终近似相等，最大化的实现了端口的均衡利用，进而提升网卡的有效利用带宽。

3.3.2　软件层面的数据包切分

鉴于在低延迟内核中已经使用了足量的QP，并发度已经足够高，而在这种情况下将部分处理步骤从硬件放到软件上并不会提升总体时延，同时还可以降低硬件的处理压力。具体而言，如图3-3所示，把一部分硬件的处理时延放到软件层面实现，确实会增加软件层面上的处理压力，即T_{WQE}增加。然而，由于低延迟内核的并发程度足够高，使用了大量的QP，这部分增加的时延完全可以被并发度数据传输时间掩盖住，如图3-4所示，进而不会影响总体时延。经测试，这个方法在硬件层面降低了每个WQE的处理时延，进而使得端到端时延降低了7.1%左右。

3.4　实验评估

本文的通信库优化的典型运行情况如下。我们的部署硬件环境为一台服务器上配备8张H20 GPU，以及8张双端口ConnectX-7 RDMA网卡。对于高带宽通信内核的实验评估，我们使用4台服务器，共32张GPU；对于低延迟的通信内核，根据不同EP规模使用不同数量的服务器。每台服务器上使用的CUDA版本为12.8。我们的对比

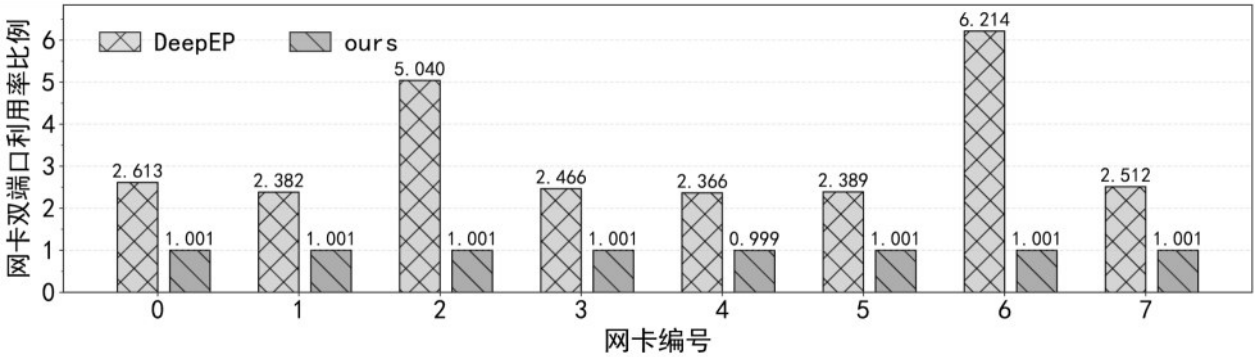

图3-9　低延迟内核端口均衡优化前后网卡端口利用率比例

基线是开源的DeepEP（当前SOTA），实验直接基于DeepEP提供的标准Benchmark进行。部署在16~32张GPU上，每个GPU实例都接收处理大量的用户请求。对于高带宽通信内核的实验评估，我们测量了每个GPU上负责的总token数不同的情况下，dispatch和combine阶段的带宽情况；对于低延迟通信内核的实验评估，我们在不同EP规模下、每张GPU上负责128个token的实验条件下对时延进行了测量。

3.4.1　高带宽通信内核的实验评估

在不同的token批次大小下，我们的高带宽内核优化版本相较之普通版本DeepEP，如图3-10所示，dispatch的带宽提升了59.45%~114.31%，如图3-11所示，combine的带宽提升了93.22%~117.55%。

3.4.2　低延迟通信内核的实验评估

在不同EP规模下，我们的低延迟内核优化版本相较之普通版本DeepEP，如图3-12所示，dispatch的时延降低39.26%~47.43%，combine的时延降低26.96%~40.49%。

3.4.3　大规模集群训练验证

为了验证方案在万卡级规模下的扩展性，我们在由8192张H800 GPU组成的星脉3.0集群上进行了私有MoE大模型的预训练实测。实验对比了采用星脉2.0多轨架构配合标准NCCL通信库的基线方案，与采用星脉3.0单轨架构配合自适应通信库的新方案。

实验结果表明，得益于单轨架构带来的跨轨跳数减少以及高带宽通信内核对拥塞的有效控制，在相同的训练任务下，星脉3.0方案的端到端训练迭代速度（Iteration Speed）提升了约15%。这一显著收益主要源于All-to-All集合通信效率的提升，有效缓解了大规模专家并行带来的“通信墙”问题，证明了该方案在超大规模集群

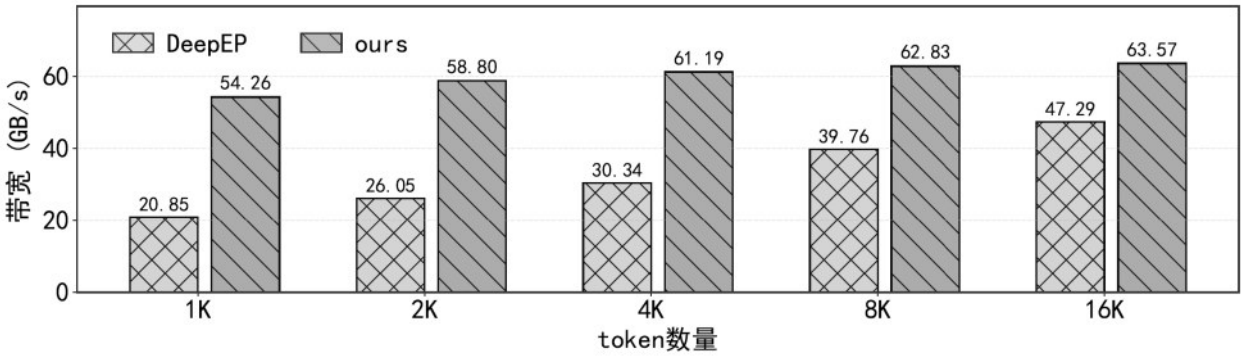

图3-10　高带宽通信内核Dispatch操作优化带宽提升

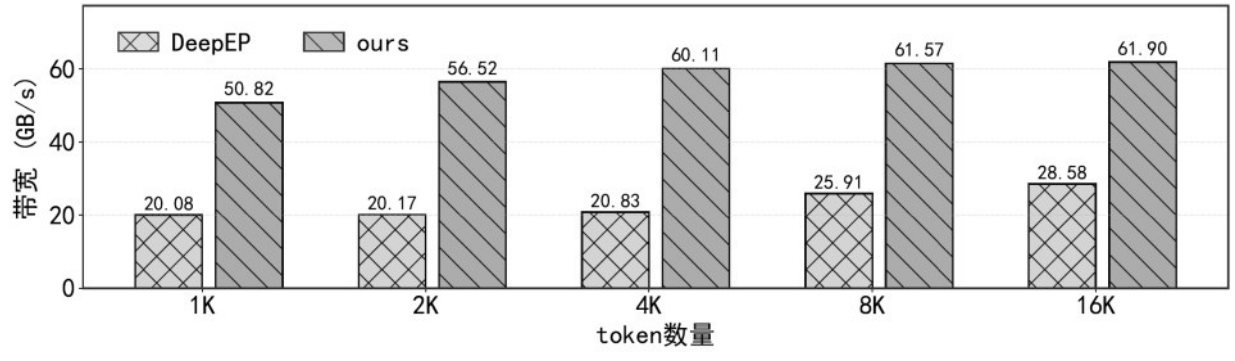

图3-11　高带宽通信内核Combine操作优化带宽提升

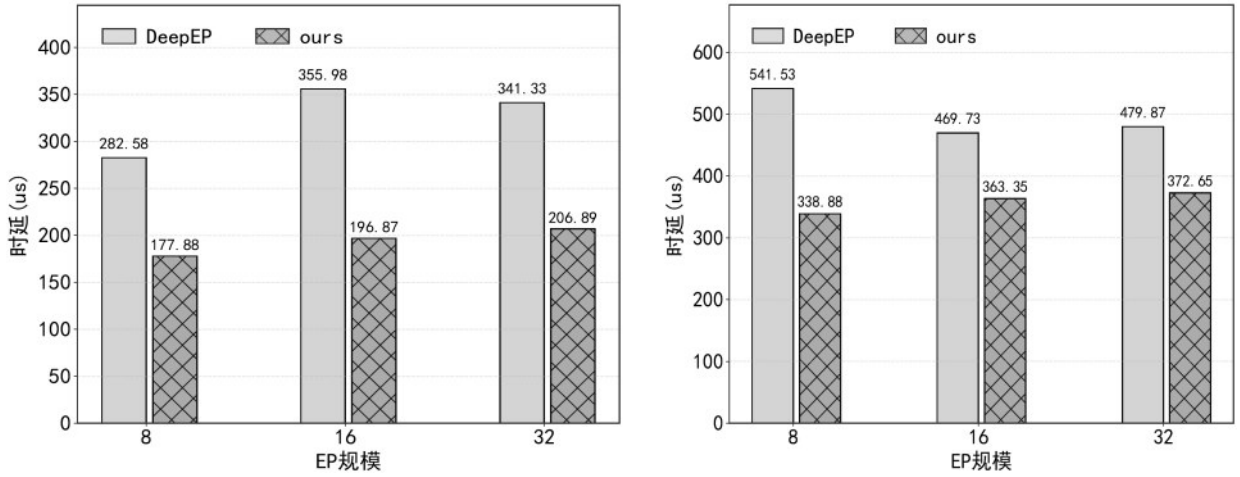

图3-12　低延迟通信内核Dispatch（左）和Combine（右）操作优化时延降低

下的工程实用价值。

4　AI Agent在网络系统运维中的应用

4.1　传统运维模式的缺陷及AI Agent的兴起

对于长周期的分布式训练任务而言，集群的稳定性决定了训练效率的上限。由于训练过程采用同步并行模式，任何单点的故障都可能导致全局训练任务暂停，迫使系统回滚至上一个 Checkpoint 进行断点续训。这种频繁的中断与恢复不仅大幅降低了有效训练时间，更造成了昂贵算力资源的巨大浪费。对于在线推理服务而言，高可用性是保障 SLA 的底线。若出现了故障不能被及时精准定位，将直接导致服务抖动甚至大面积不可用。因此，网络运维体系的效率直接决定了集群整体的训练产出率与推理服务质量。然而，面对万卡级集群的指数级规模扩张，传统运维架构已难以满足对高可靠性的严苛要求。

在传统的网络运维体系中，故障处理高度依赖于基于专家定义规则的自动化系统。面对网络规模与复杂度的指数级增长，此类静态规则系统在应对动态异常时显得力不从心，导致运维工单严重积压。大量未及时处理的工单不仅消耗了宝贵的运维资源，也对网络服务的稳定性构成了持续压力，凸显出现有运维模式在效率与可扩展性方面的不足。

深入分析表明，现有运维模式在故障定位环节存在显著短板。传统的诊断流程涉及复杂的规则匹配与人工研判，对少数核心专家的经验存在强依赖。这种强人力依赖导致故障排查路径冗长、不确定性高，致使平均故障修复时间难以有效压缩。加之专家资源的稀缺性与分布不均，极易导致关键时刻的响应延迟。因此，如何在规则系统基础上引入更智能的分析能力，以降低对专家经验的过度依赖、提升定位效率，已成为业界亟待解决的关键问题。

除定位效率外，数据层面的割裂也严重制约了运维效能的提升。在传统网络运维环境中，性能、告警、日志、历史工单以及RMA（返修记录）等关键数据分散在不同的系统中，无法统一整合后进行研判。缺乏统一的数据视图与关联分析机制，使得运维人员难以根据不完整的数据对故障进行深度分析，从而影响了根因判断的准确性与前瞻性预警能力的构建。这些挑战共同凸显了现有运维模式在数据融合与智能分析方面的局限性。

与传统运维模式的被动响应不同，AI Agent

引入了自主性、反应性与目标导向性的全新范式。借助大语言模型强大的语义理解与逻辑推理能力，AI Agent 具备了类人的自主决策与规划能力，能够主动将复杂的运维目标拆解为可执行的任务流并编排执行。特别是在处理非结构化数据方面，Agent 仅凭运维日志文本即可实现异常的自主检测、修复与预警，克服了传统 AIOps 依赖特定算法开发与模型训练的高门槛，显著降低了运维的人力成本与技术复杂度。随着企业数字化转型的深入，构建基于 AI Agent 的智能化运维体系已成为行业发展的必然趋势。

4.2 基于AI Agent的光模块故障自动诊断

4.2.1 光模块运维的批次性风险与系统建设动机

随着现代数据中心网络规模的指数级扩张，光模块作为关键的有源互连组件，其部署量级已攀升至百万级别。尽管其故障率长期维持在 5% 至 8% 的稳定区间，但庞大的基数使得故障的绝对数量显著增加。在传统运维模式下，诸如光口污染、静电放电（ESD）损伤等物理层故障往往难以被精准定位，不仅造成了运维资源的巨大浪费，也对网络的高可用性构成了严峻挑战。

尤为关键的是，工业界实践表明光模块故障呈现出显著的批次相关性。尽管同一批次产品能够通过常规的入网检测，但其潜在的工艺缺陷往往在全负荷运行阶段才会逐步显现。若沿用传统的人工被动响应模式来排查此类隐蔽的批次性风险，将耗费难以估量的时间成本与人力资源。为解决上述痛点，我们构建了“网络光模块诊断分析 AI Agent”，旨在推动光模块运维从被动响应向自动化、智能化与主动化防御的范式转变，通过数据驱动的方式实现对潜在风险的早期识别与精准治理。

4.2.2 整体设计架构

如图 4-1 所示，网络光模块诊断分析 AI Agent 采用分层架构设计，自上而下划分为用户层、大模型层与执行层。为突破单体 LLM 在复杂运维场景中存在的幻觉率高及指令遵循能力弱等局限，核心调度机制由传统的单模型驱动升级为 AI Workflow（人工智能工作流） 编排。

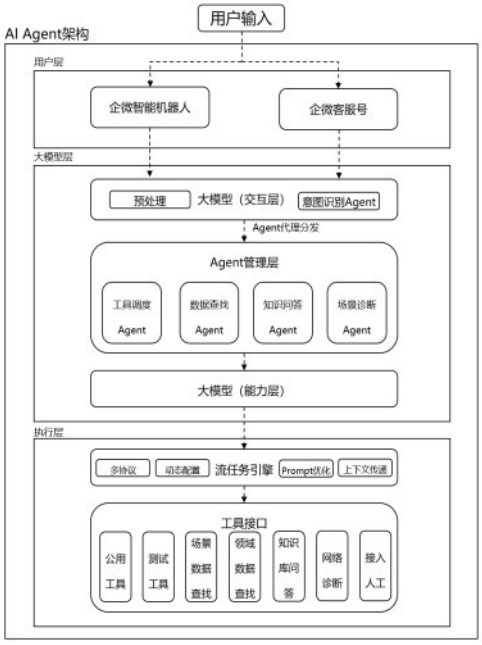

图4-1 网络光模块诊断分析AI Agent架构

AI Workflow 的核心思想是将非确定性的复杂运维任务，解耦为由多个确定性节点（Node）组成的有向无环图（DAG）。在传统的单 LLM 模式下，模型如同处理一道“开放式简答题”，其生成内容的概率空间过大，极易导致幻觉或逻辑跳跃。而在 Workflow 架构下，任务被拆解为“参数提取”、“逻辑判断”、“API 调用”等原子步骤：参数提取节点将任务转化为“填空题”，限制模型仅提取关键信息（如 IP、错误码）；逻辑判断节点将决策转化为“选择题”，限制模型在预设分支中进行路径选择。这种通过限制生成概率空间的范式转变，有效规避了模型的发散性错误，将故障诊断场景下的自动化解决率从不足 70% 提升至 80% 以上。在实际生产环境的部署中，该 AI Agent 系统已覆盖超过 10 万个光模块端口的日常巡检。数据显示，系统对光路静电损伤、脏污等常见故障的诊断准确率（Precision）

达到 96.5%，召回率（Recall）达到 98.2%。与传统人工排查相比，平均故障诊断时间（MTTD）从小时级缩短至分钟级，运维人效提升显著。

在工程实践层面，AI Workflow 采用可视化画布进行低代码编排，显著降低了业务逻辑的维护门槛。系统支持通用 HTTP 业务接口的标准化配置，无需为大模型开发大量定制适配器。此外，结合大自身的推理能力，系统可实现话术自动生成与接口缺失信息的智能追问，将流程配置的人力成本降低了约 50%，实现了开发效率与运行稳定性的双重提升。

4.2.3　主动式批次风险预警

工业界实践表明，光模块故障并非完全随机的独立事件，而是呈现出显著的批次相关性。同一厂商、型号或生产批次的光模块，往往因共性的工艺缺陷（如激光器老化加速、封装气密性不足），在特定运行周期后表现出趋同的失效特征。基于这一核心洞察，本系统设计了主动式批次风险预警机制，旨在从海量、离散的监控数据中挖掘隐蔽的共性风险。

在技术实现上，鉴于全网光模块监控指标高达千亿级的庞大规模，直接使用 LLM 进行全量分析在算力成本与响应时效上均不可行。因此，本系统采用了分层治之的混合架构，如图 4-2 所示：

1. 特征提取与算法初筛（数据层）：系统首先对海量的原始监控数据（光功率、温度、偏置电流等）进行清洗与去噪。随后，引入支持向量机（SVM）、线性回归（Linear Regression）以及专家阈值规则等轻量级传统模型，对时序数据进行特征提取与异常初筛。这一层负责“做减法”，高效过滤掉 99% 的正常数据，精准捕捉潜在的离群点与趋势异常。

2. LLM 推理与风险研判（Agent 层）：经算法初筛后的异常特征数据被整合并转化为结构化的提示词，输入给 AI Agent 的核心大模型。Agent 利用其强大的逻辑推理能力，综合考虑设备批次信息、历史故障模式及环境上下文，对异常数据进行二次研判与概率预测。LLM 能够识别出传统算法难以察觉的非线性关联，从而输出精准的异常概率与风险等级。

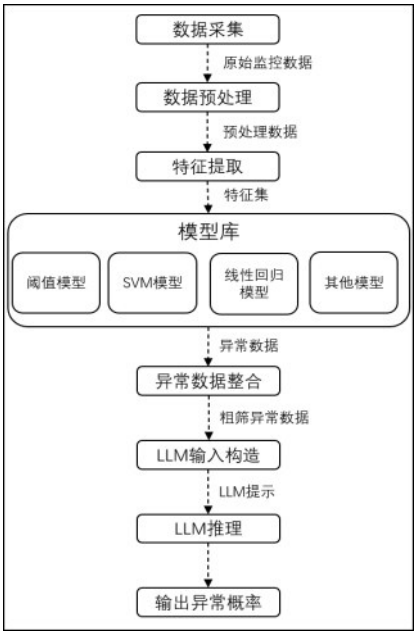

图 4-2　基于“传统算法初筛 + LLM 深度推理”的批次风险预警流程

基于上述架构，系统能够自动生成《主动式批次风险预警报告》，精准圈定高风险批次，并结合故障根因库给出“重点监控”、“固件升级”或“计划性退换”等处置建议。这一机制实现了从单点被动维修到批次主动防御的范式升级，有效避免了大规模级联故障的发生。

4.2.4　基于LLM的交互式故障诊断分析

网络光模块诊断分析 AI Agent 将诊断模式从传统的“2 个关键指标+专家规则堆”升级为“38 个细粒度指标+算法粗筛异常+LLM 精准研判”的全新架构。基于 LLM 和 RAG（检索增强生成）技术，AI Agent 提供三大核心交互能力：运营诊断、风险分析问答和知识库问答。在运营诊断方面，如图 4-3 所示，运维人员可直接使用自然语言发起请求，AI Agent 通过 LLM 解析用户意图，

并结合实时数据生成诊断报告，简化了传统繁琐的查询流程。

本系统实现了光模块诊断范式的根本性变革，将传统的“关键指标阈值 + 专家规则堆叠”模式，升级为“全量细粒度指标 + 算法初筛 + LLM 精准研判”的全新架构。基于大语言模型（LLM）与检索增强生成（RAG）技术，AI Agent 构建了运营诊断、风险探查与知识问答三大核心交互能力，彻底改变了人机协作模式：

1. 自然语言驱动的即时诊断： 如图 4-3 所示，运维人员无需掌握复杂的查询指令，即可直接通过自然语言发起诊断请求（例如：“分析 Core-Sw-01 端口的光衰情况”）。Agent 利用 LLM 强大的语义解析能力精准识别用户意图，自动调取 38 项细粒度实时数据，并生成包含根因分析的诊断报告。这一机制显著简化了传统繁琐的数据查询与关联分析流程，降低了运维操作门槛。

2. 深度的风险归因分析： 针对批次预警清单中的高风险设备，系统支持多轮问答式的深度探究。如图 4-4 所示，风险分析问答功能允许用户针对预警清单中的高风险批次进行深度探究，AI Agent 会对风险分析问题给出更加深度详细的回答。

4.2.5 智能化知识库服务

针对运维领域中知识资产碎片化、隐性经验难以复用的痛点，系统构建了智能化向量知识库服务。该服务旨在将海量离散的运维数据转化为可被机器理解与检索的结构化知识资产。

在知识构建层面，系统深度整合了海量的历史工单记录、典型故障案例库以及产品标准作业程序（SOP）。依托 TCADP 智能体平台的数据处理能力，系统对这些多源异构的非结构化数据进行了标准化的清洗与切分，并将其映射为高维向量图谱进行存储。这一过程显著提升了知识的召回准确度，实现了从人工文档维护到自动化知识

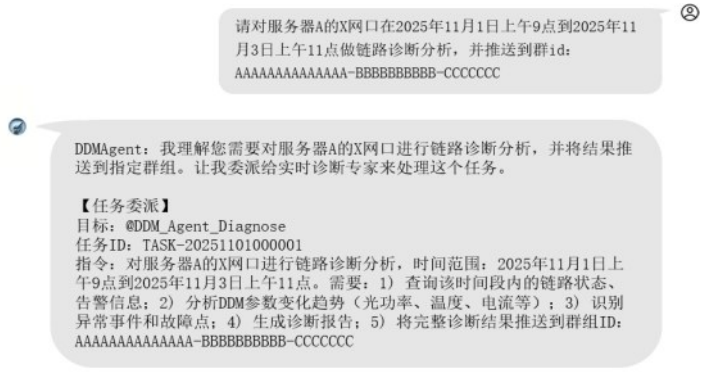

图4-3 用户使用自然语言进行交互

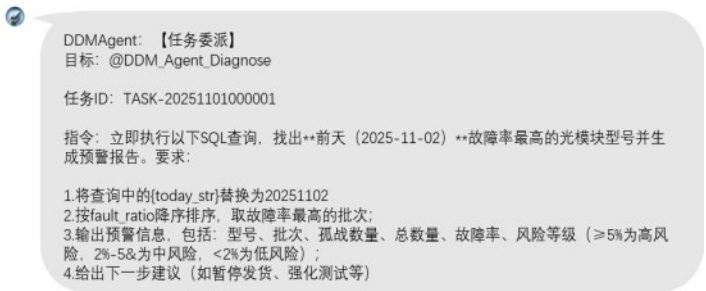

图4-4 用户要求进行深度分析

工程的范式转变。

在应用执行层面，当 AI Agent 在诊断过程中识别到特定的故障特征或语义意图时，会自动触发检索增强生成（RAG）机制。系统不仅能基于语义相似度快速定位高度相关的历史处置经验，还能综合当前的告警上下文，为一线运维人员输出具备强可执行性的下一步操作建议。这种机制有效降低了人工排查的复杂度与试错成本，辅助团队快速完成故障闭环，显著提升了运维作业的标准化水平与处置效率。

5　结束语

本文聚焦于大规模 MoE 训推一体场景中日益严峻的通信瓶颈与运维挑战，系统性地阐述了高性能网络基础设施的设计与实践。通过构建星脉 3.0 扁平化单轨网络架构，兼顾训练高带宽与推理低时延需求，有效突破了跨轨通信限制；设计了以 GPU 为中心的 All-to-All 通信内核，利用软硬协同技术同时解决了大包吞吐瓶颈与小包长尾时延问题；引入基于 AI Agent 的智能运维体系，实现了从被动响应到主动预警的运营模式升级。实验与实践表明，该全栈解决方案有效缓解了“通信墙”问题，为迈向万亿参数时代的大规模训练与在线服务提供了统一的高性能系统底座。

参考文献：

[1] 国家数据局 . 国家数据局：国内多数模型训练使用中文数据占比超 60%[EB/OL]. (2025-08-19)[2025-12-08]. https://www.gov.cn/lianbo/bumen/202508/content_7037033.htm.

[2] Tech Investments. A Niche Winner in the AI Data Center[EB/OL]. (2025-06-28)[2025-12-08]. https://www.techinvestments.io/p/a-niche-winner-in-the-ai-data-center.

[3] LIU Z, LIN Y, CAO Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 10012-10022.

[4] YUAN J, GAO H, DAI D, et al. Native sparse attention: Hardware-aligned and natively trainable sparse attention[C]// Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2025: 23078-23097.

[5] FEDUS W, ZOPH B, SHAZEER N. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity[J]. Journal of Machine Learning Research, 2022, 23(120): 1-39.

[6] DU N, HUANG Y, DAI A M, et al. Glam: Efficient scaling of language models with mixture-of-experts[C]. Proceedings of the International Conference on Machine Learning, 2022: 5547-5569.

[7] LIU A, FENG B, WANG B, et al. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model[J]. arXiv preprint arXiv:2405.04434, 2024.

[8] YANG A, YANG B, HUI B, et al. Qwen2 technical report[J]. arXiv preprint arXiv:2407.10671, 2024.

[9] JIANG A Q, SABLAYROLLES A, ROUX A, et al. Mixtral of experts[J]. arXiv preprint arXiv:2401.04088, 2024.

[10] JIN Z, WANG S, ZHU J, et al. BigMac: A Communication-Efficient Mixture-of-Experts Model Structure for Fast Training and Inference[C]. Proceedings of the AAAI Conference on Artificial Intelligence, 2025, 39(17): 17689-17698.

[11] JIN P, ZHU B, YUAN L, et al. Moe++: Accelerating mixture-of-experts methods with zero-computation experts[J]. arXiv preprint arXiv:2410.07348, 2024.

[12] HWANG C, CUI W, XIONG Y, et al. Tutel: Adaptive mixture-of-experts at scale[J]. Proceedings of Machine Learning and Systems, 2023, 5: 269-287.

[13] Meng Q, Zheng H, Zhang Z, et al. Astral: A Datacenter Infrastructure for Large Language Model Training at Scale[C]//Proceedings of the ACM SIGCOMM 2025 Conference. 2025: 609-625.

[14] Jiang Z, Lin H, Zhong Y, et al. {MegaScale}: Scaling large language model training to more than 10,000 {GPUs} [C]//21st USENIX Symposium on Networked Systems Design and Implementation (NSDI 24). 2024: 745-760.

[15] Qian K, Xi Y, Cao J, et al. Alibaba hpn: A data center network for large language model training[C]//Proceedings of the ACM SIGCOMM 2024 Conference. 2024: 691-706.

[16] Gangidi A, Miao R, Zheng S, et al. Rdma over ethernet for distributed training at meta scale[C]//Proceedings of the ACM SIGCOMM 2024 Conference. 2024: 57-70.

[作者简介]

夏寅贲（1975- ），男，博士，现任腾讯网

络首席架构师，主要研究方向高性能网络、智算互联系统。

王梓博（2000 - ），男，南京大学计算机软件新技术国家重点实验室博士生，主要研究方向为机器学习系统。

曹培睿（1993 - ），男，博士，现任南京大学计算机软件新技术国家重点实验室助理研究员，主要研究方向为光电混合数据中心网络，智算网络与系统等。

王智彬（1996 - ），男，博士，现任南京大学计算机软件新技术国家重点实验室助理研究员，主要研究方向为图计算及机器学习系统。

王亚晨（1979- ），男，硕士，现任腾讯云副总裁、腾讯网络总经理，现任中国通信学会网络专委会和边缘计算专委会副主任委员、算力网络专委会委员，主要研究方向智算网络、骨干网络、云网一体系统。