DOI: 10.7544/issn1000-1239.ps20241101
61(11): 2663−2663, 2024

大模型时代网络基础设施的机遇与挑战

过敏意
（上海交通大学计算机科学与工程系　上海　200240）

大模型训练已经成为推动人工智能（AI）和云计算发展的核心驱动力之一，其底层基础设施的性能直接影响训练效率和模型表现. 随着大模型规模的不断扩大，数据中心网络的性能需求也在持续提升，尤其是在支持大规模 AI 模型训练方面，成为未来云计算网络架构演进的重要研究对象. 阿里云团队通过深入分析大模型训练中的网络瓶颈和技术挑战，从“网络架构设计、通信感知调度、故障分析定位”3 个关键维度进行研究，提出了专为大模型训练设计的网络架构优化方案，实现了大规模训练任务的高效调度和网络传输优化，并在实际系统中取得了显著的性能提升. 文章从大模型训练的具体需求出发，探讨了如何通过创新性的网络设计提升负载均衡性、GPU 利用率和网络容错性，主要涵盖以下 3 个核心方面：

1）基于新型数据中心网络架构 HPN 的设计，包括引入一种两层、双平面（dual-plane）的网络架构，利用简单的结构来降低复杂网络执行开销大的问题；设计去堆叠双 ToR（top-of-rack）机制来解决传统单 ToR 交换机的单点失效风险；通过精准选路实现高效的 GPU 互联，解决了大规模并发训练的通信瓶颈问题.

2）基于多任务通信调度优化器 Crux 的设计，包括定义 GPU 计算密集度（GPU intensity）来精确区分训练任务，降低通信竞争带来的执行开销；提出优先处理高 GPU 计算密集度任务的方法，类似于 one-hot 编码构建的行动空间，覆盖了现有的通信调度算法；设计后备机制来保证多任务调度的正确性，提升了 GPU 利用率.

3）基于集合通信库的日志信息与训练框架关联分析的故障定位方案，包括将大模型训练特有的组件（如集合通信库日志、训练框架追踪输出等）进行横向关联，以深入理解系统瓶颈；针对特定故障时间进行分析，快速、精准地定位问题，降低了传统故障定位方法的延迟和开销.

AI 的发展不仅需要先进的算法和硬件支持，还依赖于底层数据中心网络架构的不断迭代和优化. 阿里云提出的新型数据中心网络架构走出了关键的一步，但在应对更大规模的训练任务、更多样的训练任务故障，以及推动相关技术生态系统的成熟方面，依然有很大的成长空间. 如何更好地适应未来大模型的超大规模训练需求、如何进一步加速训练与推理一体化架构的普及、如何应对多样化多层次的故障定位缝隙都是阿里云和业内同仁需要持续探索的问题. 值得庆幸的是，阿里云选择了开放合作的态度推动技术创新. 我相信，在这种开放协作的模式下，国内大模型领域将迎来更大的发展与突破.

评述专家



过敏意，教授，博士生导师. 主要研究方向为大数据、并行与分布计算、普适计算、计算机网络、软件工程、编译与程序优化等.

亮点论文

翟恩南，操佳敏，钱坤，关宇. 面向大模型时代的网络基础设施研究：挑战、阶段成果与展望 [J]. 计算机研究与发展，2024，61(11)：2664−2677. DOI: 10.7544/issn1000-1239.202440576