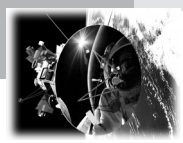

专栏

基于国产NPU的超万卡智算集群大模型训练调优实践

娄涛[1]，牛红韦华[1]，张鹏飞[2]，董江帆[3]，李攀攀[1]，
李道通[1]，许伟栋[1]，姚成辉[1]，薛连浩[1]，唐婷[1]，向洁[1]
（1. 中移（苏州）软件技术有限公司，江苏 苏州 215123；
2. 中国移动通信集团有限公司，北京 100032；
3. 中国移动通信集团设计院有限公司，北京 100080）

摘　要：为解决超万卡智算集群模型训练算效利用率低、稳定性差、调优难度高、国产技术生态差等问题，提出一种基于全国产化超万卡智算集群的大模型训练调优方案。通过自动分布式策略推荐、流水线并行优化、overlap优化和全链路profiling等技术，在16 384个国产NPU加速卡上实现了405B参数大模型的预训练，模型算力利用率（model FLOPS utilization，MFU）达到了45.13%，较基准性能提升了10%以上。同时，在模型训练全流程中构建稳定性保障机制，实现训前和训中关键指标的实时监控和训练任务秒级故障诊断。实验结果表明，提出的国产超万卡智算集群大模型训练方案能有效提升算力利用率，对未来国产智算集群建设与大模型训练有重要指导意义。

关键词：超万卡智算集群；国产NPU加速卡；模型训练调优
中图分类号：TP393
文献标志码：A
doi: 10.11959/j.issn.1000−0801.2025166

Practice of large language model training optimization based on large-scale AI cluster with more than 10 000 domestic NPU

LOU Tao[1], NIU Hongweihua[1], ZHANG Pengfei[2], DONG Jiangfan[3], LI Panpan[1], LI Daotong[1],
XU Weidong[1], YAO Chenghui[1], XUE Lianhao[1], TANG Ting[1], XIANG Jie[1]
1. China Mobile (Suzhou) Software Technology Co., Ltd., Suzhou 215123, China
2. China Mobile Communications Group Co., Ltd., Beijing 100032, China
3. China Mobile Group Design Institute Co., Ltd., Beijing 100080, China

Abstract: In order to solve the problems of low computing efficiency utilization, poor stability, high difficulty in training optimization, and imperfect domestic accelerator technology ecology in AI cluster model training with more than 10 000 NPU, a large language model training optimization solution based on a completely domestic AI cluster was proposed. Through automatic distributed strategy recommendation, pipeline parallel optimization, overlap optimi-

收稿日期：2025−03−21；修回日期：2025−06−11
通信作者：李攀攀，lipanpan@cmss.chinamobile.com

zation and full-link profiling technology, the model FLOPS utilization (MFU) reached 45.13% when training a 405B large language model on 16 384 domestic NPU, which was more than 10% higher than the baseline performance. At the same time, a set of stability assurance mechanisms was built throughout the entire large language model training process to achieve real-time monitoring of key indicators before and during model training, as well as the ability to quickly diagnose faults after training task were interrupted. The experimental results show that the large language model training optimization solution proposed can effectively improve the utilization of computing efficiency, and has important guiding significance for the future construction of domestic AI cluster and large language model training.

Key words: AI cluster with more than 10 000 cards, domestic NPU accelerator card, model training optimization

0　引言

随着人工智能（artificial intelligence，AI）技术的持续进步，大模型已成为AI领域的核心驱动力，在文本生成[1]、知识推理[2]、文生视频[3]等多个领域均表现卓越，展现出非凡的潜力。这一技术不仅在学术界引发了广泛的研究兴趣，在工业界也得到了深入应用，为智能化服务奠定了坚实的技术基石。近年来，大模型的发展势头尤为强劲，业界紧密遵循规模定律（Scaling Laws）[4]，该定律明确指出算力、算法和数据是推动大模型发展的三大关键因素。随着技术的不断变革，大模型的参数规模已经从千亿级跃升至万亿级，对底层算力的需求也随之急剧增加[5]。此外，近年来国际政治经济环境错综复杂，特别是英伟达（NVIDIA）高端AI芯片面临的禁售等限制，进一步加剧了我国在高性能计算领域所面临的挑战。因此，构建高效、稳定、绿色且易用的国产大规模智算集群，并实现基于国产超万卡智算集群的大模型高效训练，已变得至关重要。

在基于国产超万卡智算集群的大模型训练调优过程中，不仅要关注模型本身的优化，还需要深入考虑大规模计算资源的精细管理和高效利用。当前，如何高效整合国产超万卡集群的计算、存储、网络等核心技术资源，充分发挥集群加速卡的运算能力，提升模型算力利用率（model FLOPS utilization，MFU），确保模型训练能够高效且长期稳定地运行已成为重要的研究方向。

目前，业界已建立多个不同形式的万卡规模集群。本文介绍了在中国移动建设的超万卡国产智算集群上进行大模型训练调优的实践过程。该超万卡国产智算集群配备国产化网络设备，成功探索1.8万个智算卡单集群部署的规模上限，可提供高达6.9 EFLOPS的强大算力支持。针对大模型训练调优，本文进一步提出了“三阶段优化，全流程监控”的创新优化方案。首先，打造训前、训中和训后集群的一键健康检查能力，涵盖加速卡健康状态检查、网络连通性检测以及集合通信带宽测试等功能；其次，提出自动分布式策略推荐、流水线并行（pipeline parallelism，PP）优化、overlap优化等技术，通过这些技术显著提升大模型训练效率；最后，设计全链路profiling技术，通过实时监控软硬件故障，有效保障模型训练的稳定高效。本文的工作为基于国产超万卡智算集群的大模型训练调优提供了新的思路和方法。

1　研究背景

1.1　大模型发展趋势

近年来，大模型的发展势头强劲，其参数规模实现了从千亿级到万亿级的跨越性增长，大模型的综合能力得到了显著提升。自2018年BERT模型被提出后[6]，大语言模型的参数量开始呈现指数级增长，从1.1亿持续攀升至2020年GPT-3模型[7]的1 750亿，2024年Meta发布Llama 3.1-405B模型[8]，参数规模达到4 050亿，成为当时参数规模最大的开源模型。与此同时，国内智谱AI的

GLM系列模型[9]、腾讯的混元大模型[10]以及Mistral AI的Mixtral系列模型[11]等也相继问世，为大模型应用和智能化服务拓展奠定了坚实的技术基础。

然而，随着模型参数量的增长，大模型的训练效率低下问题日益凸显。庞大的参数量及训练数据集意味着需要消耗巨大的计算资源，并且训练周期大幅延长。以Llama 3.1-405B为例，其训练过程耗费数万个加速卡连续训练数月才完成。因此，构建高效稳定的用于大模型训练的超万卡集群，成为亟待解决的关键问题。

1.2 万卡智算集群建设现状

随着大模型对计算资源的需求持续提升，全球范围内智算集群的建设呈现蓬勃发展的态势[12]。在国际上，Meta、xAI和微软等国际科技巨头持续突破规模极限。Meta已建成两个分别配备24 576个图形处理单元（graphics processing unit，GPU）的集群，并计划构建一个拥有35万个H100 GPU的超大规模集群；xAI与超微电脑（SMCI）合作，成功部署了拥有10万个NVIDIA H100 GPU的Colossus集群，刷新了业界纪录；微软计划到2025年年底向OpenAI提供约30万个NVIDIA最新的GB200 GPU，以支持OpenAI在AI领域的研究。

在国内，字节跳动、摩尔线程、算丰信息等公司也纷纷建立了千卡、万卡级智算集群。字节跳动搭建了一个拥有12 288个Ampere架构GPU的训练集群，并研发了MegaScale架构[13]用于大语言模型的训练。此外，国产化智算集群建设也已进入全面发展阶段。摩尔线程打造的千卡千亿模型训练平台——摩尔线程KUAE智算中心，以国产全功能GPU为底座，能够提供软硬件一体化的完整解决方案[14]。算丰信息携手优刻得，在优刻得青浦智算中心上线国产千卡智算集群，支持千卡规模、千亿参数级别的大模型训练和推理任务[15]。总体而言，尽管我国在千卡规模智算集群建设方面取得了显著进展，但国产万卡规模的集群建设仍处于起步与探索的关键阶段，面临诸多挑战，需要进一步突破与发展。

1.3 万卡智算集群大模型高效训练挑战

随着模型参数规模的持续增长以及智算集群技术的迅猛发展，如何确保大模型高效且稳定的训练已成为当前亟待解决的关键问题，并吸引了国内外众多研究团队的深入研究。文献[16]基于Megatron-LM框架，通过消融实验优化训练配置，提升了模型训练效率。Meta使用超过16 000个NVIDIA H100 GPU来训练Llama 3.1-405B模型，结合预训练优化（上下文扩充、退火策略）和直接偏好优化（direct preference optimization，DPO）算法，显著提升了模型性能[8]。微软DeepSpeed团队提出的ZeRO++技术通过量化通信和分级权重分割优化，降低了大规模集群训练的通信开销，在4 096卡集群上仍能保持高效扩展[17]。Colossal-AI创新性地提出“6维并行技术”和Gemini内存管理系统，在大规模集群训练时能大幅度降低显存需求[18]。在国产AI芯片适配方面，华为昇腾和寒武纪等厂商取得显著突破。华为构建了从昇腾910训练芯片到Atlas集群的完整算力体系，并推出MindSpore框架及AscendSpeed加速库，支持千亿级大模型的分布式训练与推理优化[19]。寒武纪则基于思元系列芯片研发了Torch-MLU插件及BangTransformer引擎，完成对国内外主流大模型的适配，其动态架构设计显著提升了训练能效比[20]。

大模型高效训练虽然取得了一些进展，然而，当将这些结论应用于超万卡规模的国产智算集群进行大模型训练时，还面临一系列挑战，主要表现在以下几个方面。

（1）国产生态适配不完善。国产训练框架在应用程序接口（application program interface，API）设计、优化策略、权重定义等方面与业界主流框架存在较大差异，无法直接快速移植到国

产平台生态中。

（2）集群稳定性差。随着智算集群规模的扩大，软硬件故障的发生频率显著增加，整个通信网络变得更加脆弱。此外，大规模智算集群中常出现网卡闪断、慢节点慢链路等故障，对整个训练任务造成严重影响，导致模型训练效率大幅下降。

（3）训练策略调优难。PP 的引入虽然提高了训练效率，但也带来了通信时延、负载不均衡及梯度累积同步等挑战，如何合理划分模型阶段、优化通信机制并平衡负载，成为调优的关键。

（4）性能瓶颈分析复杂。在训练过程中，故障呈现类型多样、传播范围复杂和性能状态难以捕捉等特点，使得故障定界定位变得极其困难。

2　国产超万卡智算集群整体系统设计

随着业界主流大模型的快速迭代与发展，模型训练对 AI 算力资源的需求不断增加。为加快国产人工智能加速卡的技术成熟度和生态建设，构建标准化能力体系，本文从计算、存储、网络等多维度、多方面考虑，采用高度优化设计与灵活的网络架构，通过标准化、层次化、模块化的策略实现了资源的高效整合与灵活扩展。

2.1　计算：高速互联，提升算力

本文所述的国产超万卡智算集群共建设国产化智算服务器 2 304 台，配置卡数达 18 432 个，提供约 6.9 EFLOPS 算力，能够支持深度学习、推理等多种 AI 计算任务，可大幅提升超大参数模型的训练速度。

2.2　网络：自研组网架构，提升线性加速比

本文所述的国产超万卡智算集群采用高度优化与灵活的网络架构，通过分平面设计策略实现管理面、业务/存储面、参数面、数据面的 4 平面解耦，确保整个智算集群系统的安全性、可扩展性及高效运行能力。国产超万卡智算集群组网架构如图 1 所示。其中，管理面用于全网资源的统一运维管理，以实现集群高效运维

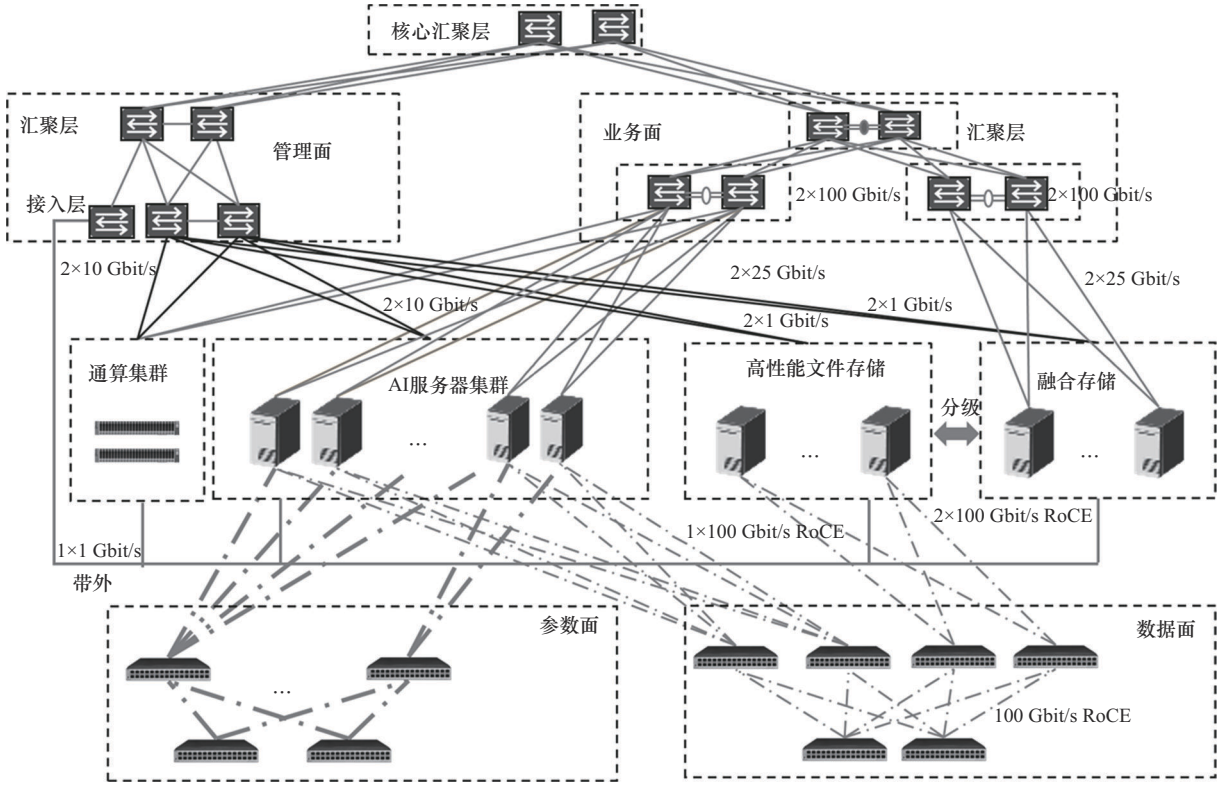

图 1　国产超万卡智算集群组网架构

和监控检测；业务/存储面用于用户访问智算集群，并通过该平面上传训练数据以及导出训练模型；参数面采用基于200 Gbit/s以太网的远程直接存储器访问（RDMA over converged Ethernet，RoCE）技术组网实现高速无损转发，用于模型训练过程中的卡间通信，并通过全调度以太网GSE 1.0逐流负载均衡方案，实现训练任务卡间通信路径调优；数据面采用100 Gbit/s RoCE网络组网，用于GPU集群访问高性能文件存储。

2.3 存储：融合技术，释放先进存力

为实现数据归一无冗余、训练流程无等待，以节约存储容量，提升训练效率，本文所述的国产超万卡智算集群存储系统采用文件、对象融合协议存储方案，同时为支持冷热数据自动分级存储，数据在高性能存储与普通存储间自动流动，节省输入/输出（input/output，I/O）访问开销，引入数据自动分级功能。存储系统融合分级架构如图2所示。

在智算服务器端，采用独立的100 Gbit/s RoCE高速数据面，并在安装可移植操作系统接口（portable operating system interface of UNIX，POSIX）并行客户端的情况下，存储集群系统的整体纯读写性能分别可达2 TB/s和1.1 TB/s。读写混合场景下的高性能文件存储单集群性能指标见表1，这样的配置可有效提升超万卡智算集群中模型训练过程中checkpoint的保存及加载效率。

表1 高性能文件存储单集群性能指标

指标名称	指标数值
单集群存储容量	12 PB
读写比6:4带宽	1 250 GB/s
读写比6:4大文件8KFLOPS	1 000万
时延	<10 ms

本文所述国产超万卡智算集群整体依托移动云智算云操作系统底座，一体化整合纳管计算、网络、存储智算资源，打造弹性、灵活的统一池化管理与调度能力；针对多租户安全隔离需求，实现智算裸金属远程直接存储器访问（remote direct memory access，RDMA）参数网络信息自动化配置及多租户训练业务流量隔离。

3 大模型训练调优方案

当前，大规模模型训练面临诸多挑战，主要包括模型训练性能提升、集群资源利用率优化以及快速故障定位等方面。本文深入研究了大规模模型高效训练的优化方法，基于405B模型和自有预训练数据集，提出了基于国产超万卡智算集群的大模型训练解决方案，有效提升了集群MFU。

3.1 模型国产化适配

模型国产化适配是为了将大模型从GPU平台迁移至国产平台，本文提出了一套模型迁移工具，包括模型迁移分析、代码迁移、权重迁移和迁移验证等核心步骤。基于本文所提出的迁移工具，已完成对40多款主流开源大模型在国产智算

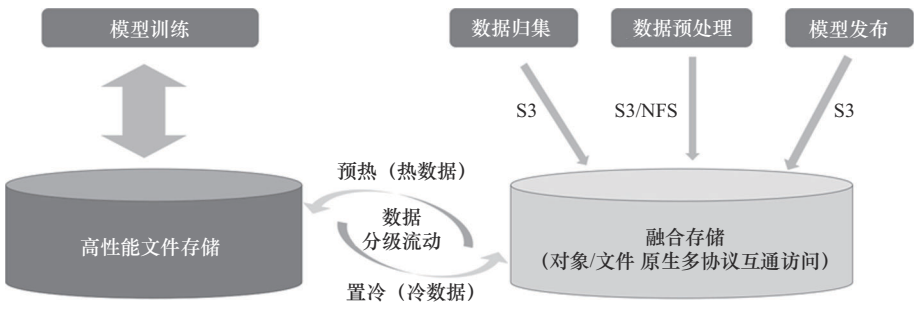

图2 存储系统融合分级架构

集群上的迁移适配验证。

迁移分析是模型国产化适配的首要阶段，本文深入剖析了目标模型的网络架构、设计理念和开源实现，对模型代码中涉及的API、算子及第三方库在国产平台上的兼容情况进行了全面评估，并精准分析出需要进行适配的部分。

代码迁移阶段主要包含训练框架适配、分布式框架适配和训练代码适配这3个方面。本文将原模型从PyTorch框架迁移到国产平台框架，并且针对不兼容的API进行了重新编写。在分布式训练策略方面，本文针对数据并行（data parallelism，DP）、张量并行（tensor parallelism，TP）和PP等技术，对Megatron-LM等分布式框架进行了适配性开发，以确保其在国产平台上的高效运行。此外，本文对训练代码中的启动脚本、数据读取、优化器及学习率等模块进行了校验与调整，以确保模型在国产平台上的正确执行。

权重迁移方面，本文将原有GPU平台训练所得的模型权重转换为国产平台支持的格式，并针对不同框架和分布式策略进行适配开发，有效保障了迁移后模型结构与精度的一致性。

本文通过loss一致性对齐进行了严格的迁移验证，loss对齐可以确保模型优化过程保持一致，避免因优化器或超参数的差异导致训练效果不稳定。本文通过计算迁移前后loss曲线的余弦相似度进行验证，结果表明，迁移后的模型在国产平台上能够保持与原平台一致的精度。

3.2　高效模型训练

大语言模型由于参数量大，常采用分布式方法将任务分配到多个计算节点来加速模型训练，然而，这也给计算资源分配、通信开销控制和负载平衡等带来了挑战。模型调优旨在通过分布式策略配置优化等方法来提升模型训练效率，加强智算集群MFU。为实现模型的高效训练，本文通过自动分布式策略推荐、流水线并行优化、overlap优化、全链路profiling监控和集群稳态检查等技术来提升大模型训练效率。

3.2.1　自动分布式策略推荐

在大模型分布式训练中，可配置的参数项逐渐增多，包括DP、TP、PP以及批次大小等。内存与性能受各种配置的影响，人工调优难度大，且效率低。基于上述挑战，本文设计了一种基于回归算法的并行策略自动搜索算法。该算法能够在给定模型结构和集群配置的条件下，基于算法与专家经验，自动推荐性能较优的并行策略，有效缩短模型调优时间。并行策略自动搜索算法流程如图3所示，该算法通过对不同模型和配置进行模拟运行（即模拟训练过程，而非实际开展训练），采集了大量模型训练数据，在此基础上，通过数据特征提取和筛选，结合回归算法构建内存预估模型。并行策略自动搜索算法的主要流程如下。

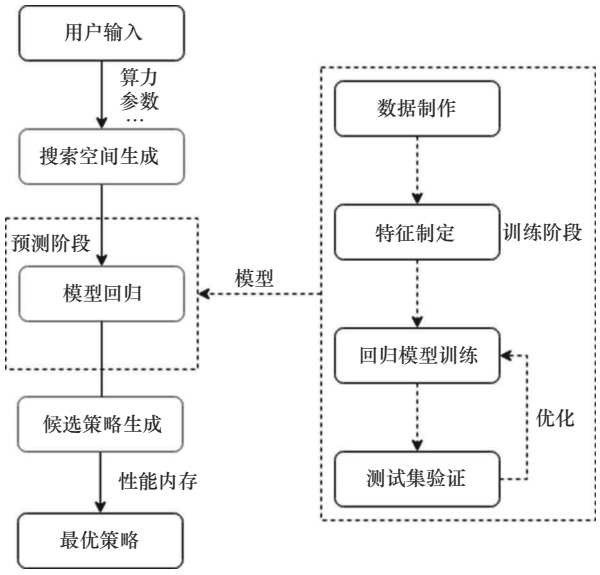

图3　并行策略自动搜索算法流程

（1）根据模型结构和集群信息的约束条件，生成训练配置的搜索空间。

（2）利用内存模型对搜索空间中的配置进行建模和预测，以排除可能导致内存溢出的配置。

（3）结合专家经验，筛选出性能较优的并行配置策略。

并行策略搜索与专家调优耗时对比见表2。由表2可知，采用并行策略自动搜索算法后，并行策略配置速度提升了83%，有效提升了模型调优效率。

表2　并行策略搜索与专家调优耗时对比

方案	耗时
专家调优	72 h
并行策略搜索+专家调优	12 h

3.2.2　PP优化

传统的一前一后（1 forward 1 backward，1F1B）PP策略[21]的思想是每进行一次前向运算，就做一次后向运算，虽能有效分割模型，减少部分空泡，但在运行过程中仍存在较高的空泡率，影响了整体性能。1F1B流水线并行策略如图4所示。

为了解决这一问题，本文使用了流水线重调度技术，在不增加设备数量的前提下，将模型训练中的前向传播和反向传播过程进一步细分stage，通过少量通信量的增加，实现更低的空泡比率。流水线重调度算法如图5所示。

此外，在PP场景中，第一个stage包含嵌入层，使得显存成为瓶颈，而最后一个stage需要计算输出和损失，使得计算成为瓶颈。本文通过自适应的流水切分，动态调整首尾stage的容量，实现了显存和计算的均衡，保证了模型训练的稳定。

3.2.3　overlap优化

在大语言模型训练中，高效利用计算资源至关重要。overlap技术通过并行执行模型训练中的计算与通信进程，旨在加速执行过程并缩减等待时间。大模型训练在进行梯度更新时，DP组的通信需要在反向计算完成后进行，这种顺序执行流程会导致计算与通信之间存在一定的资源空置时间。在PP中计算和通信也存在串行过程，从而导致资源利用率降低。

为了解决上述问题，本文提出一种全新的overlap优化方法，如图6所示。该方法融合了PP和DP，实现了计算与通信的重叠执行，通过将反向传播中梯度计算和梯度通信过程并行、将流水线之间的前向传播和send或receive操作并行，使得计算与通信任务能够在不同阶段同步执行，

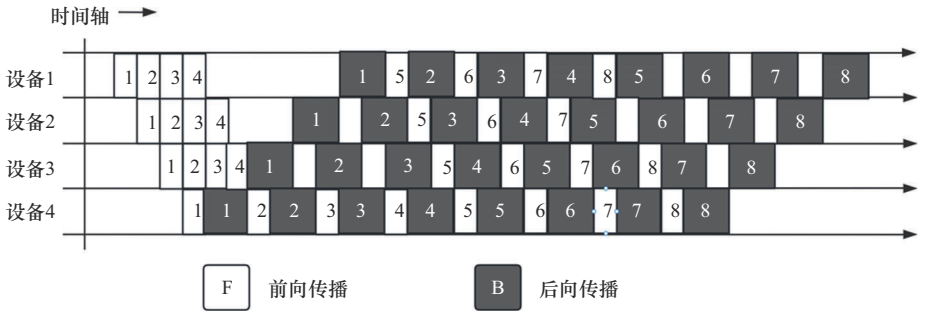

图4　1F1B流水线并行策略

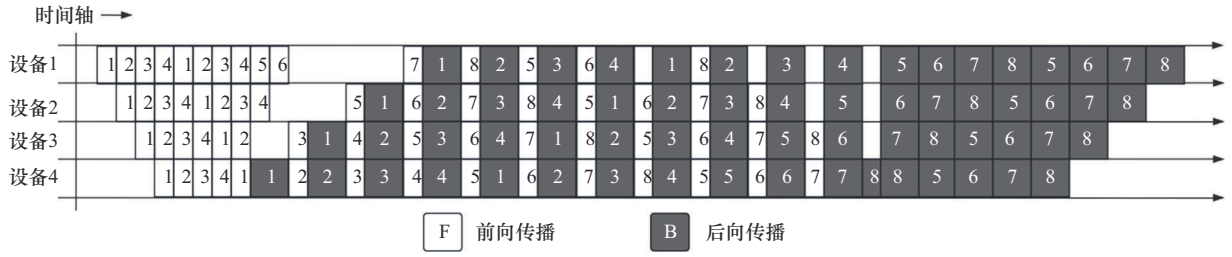

图5　流水线重调度算法

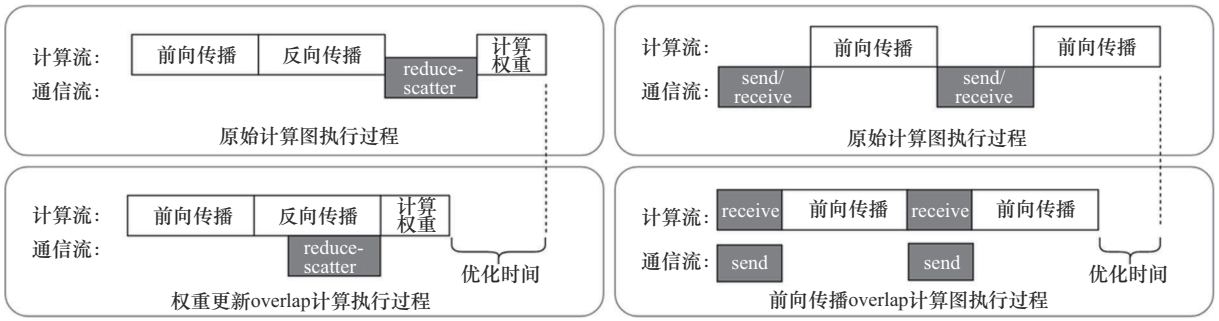

图6　overlap优化方法

大幅减少了训练过程中的空闲时段，进一步提升了整体的计算效率。相较于传统的PP和DP方法，在相同的硬件配置下，本文的overlap机制可显著提高大模型的训练效率。

3.2.4　全链路profiling监控

在大模型超万卡集群分布式训练过程中，硬件故障难以避免，且故障定位较为困难。传统的profiling方法在进行集群性能分析时，存在在线解析耗时过长、离线解析不够及时以及采集步数较少等问题，导致无法实时采集和分析集群性能。为了应对上述挑战，本文设计并提出了一种全链路profiling方法，该方法通过采集和分析全链路模型训练数据，具备动态profiling、集群瓶颈分析、模型性能跟踪、数据可视化等功能。

全链路profiling架构如图7所示，通过在框架中进行三级埋点，实现profiling数据动态采集。首先，在通信链路中设置埋点，收集通信数据，用于分析通信过程中的信息交互情况；其次，对模型训练损失进行埋点，实时监控模型损失变化，评估模型训练的效果与稳定性；最后，在系统处理环节埋点，获取系统数据，全面掌控集群在训练过程中的运行状态。通过三级埋点，本文构建了一个覆盖全链路生命周期的监控网络，能够实时跟踪模型训练过程。

在模型训练任务中，采集到的数据将上报至集群监控进程，经过专门的数据处理和格式化操作后，进行可视化展示，以实现集群瓶颈分析、节点剔除和故障定位等功能，为维护集群的稳定性提供了强有力的支持，保障了模型的高效稳定训练。

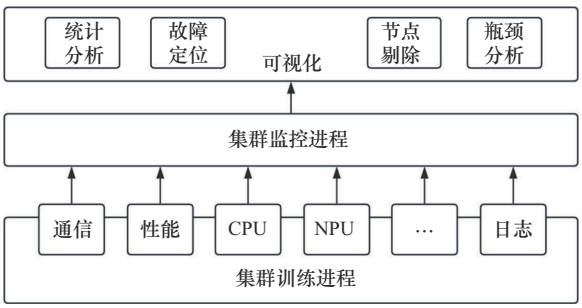

图7　全链路profiling架构

3.2.5　集群稳态检查

随着集群规模的扩大，硬件故障的发生频率显著增加，集群的稳定性成为衡量系统性能的重要指标之一。为此，本文构建了集群稳态检查能力，旨在通过全面的检查和监控，及时发现并解决可能影响集群稳定性的问题，从而保障系统的持续高效运行。

集群稳态检查系统架构如图8所示，该系统主要包括数据采集层、数据处理层和应用展示层。数据采集层利用集群命令工具收集相应数据；数据处理层对采集的数据进行处理和统计分析；应用展示层在训前、训中和训后3个阶段对集群进行稳态检查。具体来说，在训前阶段，对集群芯片温度、算力性能和内存占用等指标进行监控，以确保训练环境达标；在训中阶段，通过实时监控设备告警和链路闪断/拥塞来保障

训练过程的稳定性和连续性；在训后阶段，通过排查残留进程、清理无效文件和确保芯片健康度等步骤，确保集群环境的稳定。本文提出的集群稳态检查系统实现了对集群稳定性的全面监控和管理，有效提高了智算集群的稳定性。

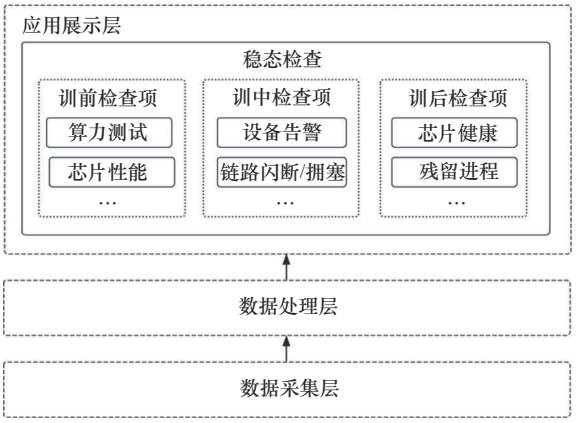

图8　集群稳态检查系统架构

4　实验结果及分析

本节主要介绍在国产超万卡智算集群环境下，针对405B模型进行训练调优后的性能结果，并深入分析各项优化技术的实际效果。

4.1　模型训练结果

为评估针对国产超万卡智算集群所提出的模型训练优化策略的有效性，本文在保持模型训练计算资源不变的情况下，对比分析所提多种优化技术与基线实验条件下集群MFU的变化情况。

基线实验中的配置：TP设为16，PP设为8，微批次大小（micro batch size，MBS）设为1，全局批次大小（global batch size，GBS）设为512，在此配置下测得的基线MFU为34.57%，对应算力为130 TFLOPS。模型在不同策略下训练的MFU（16 384 NPU）见表3。表3展示了在16 384个加速卡上进行405B模型预训练时，采用不同优化技术下的MFU变化情况。

表3　模型在不同策略下训练的MFU（16 384 NPU）

调优策略	MFU
基线	34.57%
调整并行策略TP=8、PP=16	41.06%
增大MBS=2，增大GBS=65 536	42.58%
优化PP切分策略	43.88%
优化DP+PP通信重叠	45.13%

实验结果表明，通过调整模型并行策略和批次大小、优化PP切分策略以及优化PP和DP的通信重叠，本文实现了MFU的显著提升。具体来说，通过调整并行策略TP和PP数值，使MFU提升了6.49%；基于全链路profiling技术，发现加速卡显存占用率存在瓶颈，通过调整MBS和GBS实现了1.52%的MFU提升；优化PP切分策略降低了单个PP组内首尾加速卡的显存占用，使MFU提升了1.3%；通过优化PP和DP间的通信重叠减少了通信开销，使MFU提升了1.25%。综上所述，405B参数模型训练MFU相较于基线性能提升了10.56%，本文提出的优化方法在提高大模型训练MFU方面具有显著效果。

4.2　集群MFU对比

本文基于国产超万卡智算集群进行模型训练调优实践，在405B参数规模的模型训练中实现了45.13%的MFU，显著优于现有主流方案。不同方案下的训练调优结果见表4。文献[22]指出，DeepSpeed框架结合ZeRO-1和MiCS优化，在1 024个A100 GPU上训练InternLM 7B时，MFU约为36%。文献[13]基于Megatron框架分别在12 288个和8 192个NVIDIA GPU上训练175B模型，MFU分别达到了41.2%和43.3%。本文方案在更大的硬件规模和模型参数量下，仍实现了9.13%、3.93%和1.83%的MFU提升。文献[8]采用基于Megatron框架的4D并行方案，并融合了部分DeepSpeed的设计思想，在16 384个NVIDIA GPU上训练Llama 3.1-405B模型，MFU达到了38%。本文方案在同等硬件规模和模型参

数量下，训练效率提升了 7.13%。针对国产智算集群，华为采用基于昇腾 NPU 集群的自研混合并行框架，在盘古 Ultra 135B 参数模型上实现了基线 MFU 43%[23]，较本文方案低 2.13%。

表 4　不同方案下的训练调优结果

方案	加速卡数量/个	模型参数量	MFU
DeepSpeed[22]	1 024	7B	36.00%
Megatron+DeepSpeed[8]	16 384	405B	38.00%
Megatron[13]	12 288	175B	41.20%
华为自研框架[23]	8 192	135B	43.00%
Megatron[13]	8 192	175B	43.30%
本文方案	16 384	405B	45.13%

实验结果表明，本文通过自动分布式策略推荐、PP 优化、overlap 优化、全链路 profiling 监控和集群稳态检查等一系列技术，不仅验证了其在国产化环境下的适用性，且有效提升了国产化超大规模模型训练的硬件利用率。

5　结束语

本文提出一种国产超万卡智算集群大模型训练调优方案，有效解决了超万卡智算集群模型训练算效利用率低、稳定性差、调优难度高和国产技术能力弱等问题。本文提出自动分布式策略推荐、PP 优化、overlap 优化、全链路 profiling 监控和集群稳态检查等方法，有效提高了模型训练性能和稳定性，在 16 384 个国产加速芯片上，实现 405B 大模型预训练 MFU 达 45.13%，较基准性能提升了 10.56%。

未来的研究，将持续聚焦于超大规模集群下大模型训练的自动化性能调优，探索超万卡集群训练万亿参数量大模型的性能上限，同时也将致力于模型训练全链路监控的研究，通过自动故障诊断系统解决超万卡集群训练任务故障定位困难的问题，以保障集群的稳定高效训练。本文的工作将为国产化智算集群建设和大模型训练研究提供经验借鉴和指导。

参考文献：

[1] LI J Y, TANG T Y, ZHAO W X, et al. Pre-trained language models for text generation: a survey[J]. ACM Computing Surveys, 2024, 56(9): 1-39.

[2] SINGHAL K, AZIZI S, TU T, et al. Large language models encode clinical knowledge[J]. Nature, 2023, 620(7972): 172-180.

[3] ANNEPAKA Y, PAKRAY P. Large language models: a survey of their development, capabilities, and applications[J]. Knowledge and Information Systems, 2025, 67(3): 2967-3022.

[4] KAPLAN J, MCCANDLISH S, HENIGHAN T, et al. Scaling laws for neural language models[J]. arXiv preprint, 2020: 2001.08361.

[5] 中国软件评测中心 . 人工智能大语言模型技术发展研究报告（2024 年）[R]. 2024.
China Software Testing Center. Research report on the development of artificial intelligence large language model technology (2024)[R]. 2024.

[6] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint, 2019: 1810.04805.

[7] BROWN T B, MANN B, RYDER N, et al. Language models are few-shot learners[J]. arXiv preprint, 2020: 2005.14165.

[8] GRATTAFIORI A, DUBEY A, JAUKHRI A, et al. The Llama 3 herd of models[J]. arXiv preprint, 2024: 2407.21783.

[9] DU Z X, QIAN Y J, LIU X, et al. GLM: general language model pretraining with autoregressive blank infilling[J]. arXiv preprint, 2021: 2103.10360.

[10] 腾讯云 . Tencent Hunyuan-Large: 开源业界参数规模最大、效果最好的 Transformer 结构的 MoE 模型[EB]. 2023.
Tencent Cloud. Tencent Hunyuan-Large: the largest-scale parameter and best-performing open-source Transformer-based MoE model in the industry[EB]. 2023.

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

[12] 通信产业网 . 万卡集群: 为什么？是什么？怎么建？[EB]. 2024.
Communications Industry Network. 10 000 GPU cluster: Why? What is it? How to build it?[EB]. 2024.

[13] JIANG Z, LIN H, ZHONG Y, et al. MegaScale: scaling large language model training to more than 10 000 GPUs[J]. arXiv preprint, 2024: 2402.15627.

[14] 摩尔线程 . KUAE 千卡智算中心解决方案[EB]. 2023.
Moore Threads. KUAE 1 000 GPUs AI center solution[EB]. 2023.

[15] 东方网 . 优刻得首个国产千卡智算集群落地，支持智源千亿

大模型训练[EB]. 2024.
Eastday.com. Ucloud’s first domestic 1 000 GPUs AI cluster lands, supporting the training of Zhiyuan’s 100-billion-parameter large model[EB]. 2024.
[16] HAGEMANN J, WEINBACH S, DOBLER K, et al. Efficient parallelization layouts for large-scale distributed model training[J]. arXiv preprint, 2023: 2311.05610.
[17] WANG G H, QIN H Y, JACOBS S A, et al. ZeRO++: extremely efficient collective communication for giant model training[J]. arXiv preprint, 2023: 2306.10209.
[18] LI S G, LIU H X, BIAN Z D, et al. Colossal-AI: a unified deep learning system for large-scale parallel training[C]//Proceedings of the 52nd International Conference on Parallel Processing. New York: ACM Press, 2023: 766-775.
[19] 华为, 中国信通院. 昇腾计算产业发展白皮书[R]. 2024.
Huawei, China Academy of Information and Communications Technology. White paper on the development of Ascend computing industry[R]. 2024.
[20] 寒武纪开发者社区. 思元MLU370-X8智能加速卡产品手册[EB]. 2021.
Cambricon Developer Community. Siyuan MLU370-X8 smart accelerator card product manual[EB]. 2021.
[21] HARLAP A, NARAYANAN D, PHANISHAYEE A, et al. PipeDream: fast and efficient pipeline parallel DNN training[J]. arXiv preprint, 2018: 1806.03377.
[22] CAI Z, CAO M S, CHEN H J, et al. InternLM2 technical report[J]. arXiv preprint, 2024: 2403.17297.
[23] PanguTeam, Huawei. Pangu ultra: pushing the limits of dense large language models on ascend NPUs[J]. arXiv preprint, 2025: 2504.07866.

[作者简介]

娄涛（1978– ），男，中移（苏州）软件技术有限公司副总经理，主要研究方向为云计算、智算中心、网信安全等。

牛红韦华（1989– ），女，中移（苏州）软件技术有限公司计划建设部副总经理，主要研究方向为大规模智算中心、公有云资源池架构、云管理平台等。

张鹏飞（1977– ），男，中国移动通信集团有限公司高级工程师，主要研究方向为IT信创技术、云计算及人工智能技术等。

董江帆（1987– ），男，中国移动通信集团设计院有限公司工程师，主要研究方向为云计算、智算、超算等。

李攀攀（1990– ），男，现就职于中移（苏州）软件技术有限公司，主要研究方向为云计算、智算中心、人工智能技术等。

李道通（1990– ），男，现就职于中移（苏州）软件技术有限公司，主要研究方向为公有云资源池、智算资源池方案等。

许伟栋（1993– ），男，现就职于中移（苏州）软件技术有限公司，主要研究方向为人工智能算法、大模型技术等。

姚成辉（1994– ），男，现就职于中移（苏州）软件技术有限公司，主要研究方向为人工智能算法、大模型技术等。

薛连浩（1997– ），男，现就职于中移（苏州）软件技术有限公司，主要研究方向为人工智能算法、大模型技术等。

唐婷（1997– ），女，现就职于中移（苏州）软件技术有限公司，主要研究方向为人工智能算法、大模型技术等。

向洁（1990– ），女，现就职于中移（苏州）软件技术有限公司，主要研究方向为大数据、智算中心等。