云化的智算中心万卡集群创新与实践

丁宏庆[1]，张鹏飞[1]，牛红韦华[2]，李志勇[3]，周丹媛[1]，丁国强[4]，李攀攀[2]，李道通[2]，张久仙[2]
（1. 中国移动通信集团有限公司，北京 100032；
2. 中移（苏州）软件技术有限公司，江苏 苏州 215123；
3. 中国移动通信集团浙江有限公司，浙江 杭州 311103；
4. 中国移动通信集团设计院有限公司，北京 100080）

摘　要：为解决智算中心超大规模算力集群算力可用率低、国产技术成熟度低、大规模组网效率存在瓶颈、运营运维复杂等问题，提出了一种基于云计算技术构建智算中心万卡集群的系统。采用18 432块神经网络处理单元（neural processing unit，NPU）卡和优化后的基于以太网的远程直接内存访问（remote direct memory access，RDMA）网络构建云化的智算中心万卡集群，结合软件定义网络（software defined network，SDN）技术实现RDMA网络租户隔离，实现了链路负载均衡误差小于10%，集群All-Reduce带宽达35 GB/s以上。采用优化后的分布式存储协议，实现模型断点恢复时长缩短为原来的1/2。验证结果表明，经过软硬件协同优化，国产化的NPU万卡集群不仅能够满足千亿参数大模型训练的需求，未来更可以支撑万亿参数大模型训练任务。

关键词：超级计算集群；智算中心；万卡集群；人工智能
中图分类号：TP338
文献标志码：A
doi: 10.11959/j.issn.1000−0801.2024262

Cloud-based intelligent computing center ten-thousand card cluster innovation and practice

DING Hongqing[1], ZHANG Pengfei[1], NIU Hongweihua[2], LI Zhiyong[3], ZHOU Danyuan[1], DING Guoqiang[4], LI Panpan[2], LI Daotong[2], ZHANG Jiuxian[2]
1. China Mobile Communications Group Co., Ltd., Beijing 100032, China
2. China Mobile (Suzhou) Software Technology Co., Ltd., Suzhou 215123, China
3. China Mobile Communications Group Zhejiang Co., Ltd., Hangzhou 311103, China
4. China Mobile Group Design Institute Co., Ltd., Beijing 100080, China

Abstract: To address issues such as low availability of computing power in ultra-large scale computing clusters of intelligent computing centers, low maturity of domestically produced technologies, bottlenecks in large-scale networking efficiency, and complex operations and maintenance, a system based on cloud computing technology for constructing a ten-thousand card cluster in an intelligent computing center was proposed. A ten-thousand card cluster was con-

收稿日期：2024−11−07；修回日期：2024−12−02
通信作者：张久仙，zhangjiuxian@cmss.chinamobile.com

structed using 18 432 NPU units and an optimized RDMA network. A multi-plane network architecture was adopted, in conjunction with SDN technology to achieve RDMA network tenant isolation. The network load balancing strategy was optimized, resulting in a link load balancing error of less than 10% and an All-Reduce bandwidth of over 35 GB/s. By employing the optimized distributed storage protocol, the model's breakpoint recovery time was reduced to half of its original duration. The validation results demonstrate that the domestic NPU ten-thousand card cluster, with the collaborative optimization of software and hardware, can not only meet the training needs of large models with hundreds of billions of parameters but also support the training tasks of large models with trillions of parameters.

Key words: supercomputer cluster, intelligent computing center, ten-thousand card cluster, artificial intelligence

0 引言

随着生成式人工智能在全球兴起，各国对智能计算的投入越来越高。生成式人工智能是一种基于大规模预训练模型和生成对抗网络的人工智能技术，一直以来，大模型的训练对算力的需求异常庞大，需要巨大的能耗和算力投入[1]。据不完全统计，国内已发布的大模型数量已超过100个。大模型正从千亿参数向万亿、十万亿级别参数量发展，对算力需求呈井喷式增长[2]。中国信息通信研究院发布的《中国算力发展指数白皮书（2022年）》指出，到2030年，全球算力规模预计将达到56 ZFlops，年均增速预计为65%，其中智能算力达到52.5 ZFlops，平均年增速超过80%[3]。

目前，业界已经建立了多个不同形式的万卡规模集群，但多数采用图形处理器（graphics processing unit，GPU）架构的国外技术构建。国产技术在大规模组网效率、网络隔离与负载均衡能力、软件生态成熟度和软硬件协同优化等方面还有待进一步提升。本文介绍了中国移动构建的云化智算中心万卡集群系统和创新实践，该系统采用国内NPU架构的人工智能加速卡，同时本文研究了不同网络平面的组网技术，致力于解决超大规模组网接入、网络隔离、负载均衡、任务调度和运维监控等难题[4]，旨在实现更强大的算力密度和更高效的底层计算能力，以确保万卡集群的高效性和高可用性。

1 研究背景

1.1 大模型发展趋势

在人工智能技术的飞速发展中，大型机器学习模型的涌现速度令人瞩目。特别是自ChatGPT发布以来，各种大模型如雨后春笋般涌现。模型的性能依赖于模型的规模，具体包括参数数量、数据集大小和计算量[5]。这些模型的巨大规模赋予了它们强大的表达能力和学习能力。强大的计算资源推动了计算新范式的加速实现。训练这些大模型通常需要成千上万个人工智能加速卡以及大量的时间[6]。当模型的训练参数量突破一定规模时，它们会展现出之前小模型所不具备的复杂能力和特性，类似于人类的思维和智能。大模型在自然语言处理、计算机视觉、语音识别和推荐系统等领域都有广泛的应用。2021年阿里发布全球最大AI预训练模型M6，该模型拥有多模态、多任务能力，尤其擅长设计、写作、问答，在电商、制造业、文学艺术、科学研究等领域有广泛应用前景[7]。

1.2 万卡集群建设现状

为应对智能算力需求量的激增，业界正在加快超大算力集群的建设。微软、亚马逊（AWS）、Meta、谷歌，以及国内的腾讯、百度、字节跳动、华为云、科大讯飞等都在积极推进[8]。例如，谷歌推出的超级计算机“A3 Virtual Machines”拥有2.6万块英伟达H100 GPU，并基于自研芯片搭建了TPU v5p的8 960卡集群。Meta

在2022年推出了人工智能（artificial intelligence，AI）研究超级集群（research super cluster，RSC），拥有1.6万块英伟达A100 GPU，并在2024年年初推出了两个24 576块英伟达H100 GPU集群，用于支持下一代生成式AI模型的训练。国内方面，科大讯飞在2023年建成了首个支持万亿参数大模型训练的万卡集群算力平台“飞星一号”；字节跳动也在2023年9月完成了12 288卡Ampere架构训练集群MegaScale，并启动了新的万卡集群建设。通信运营商、头部互联网公司、大型AI研发企业、AI初创企业等都在超万卡集群的建设和使用领域不断发展[9]。

1.3　万卡集群建设挑战

万卡集群建设虽然取得了一定的进展，但仍面临诸多挑战，尤其是在算力、存储、网络、基础设施以及运维等方面。

- 算力挑战：人工智能加速芯片市场由国外公司主导，国产人工智能加速卡在技术成熟度和生态建设方面与国际巨头相比仍存在较大差距。因此，如何提高由成千上万块人工智能加速卡组成的集群的算力使用效率，是一个挑战，这需要采用系统工程方法，并进行软硬件全栈整合优化[10]。
- 存储挑战：智算中心的存储系统承载着大模型训练的所有数据，传统存储解决方案存在跨池拷贝数据导致的效率低下问题。这需要建设易共享、高性能、易扩展的统一数据底座。
- 网络挑战：集群规模的快速增长对网络互联技术提出了更高的要求，这需要超大容量的网络，并探索新的大规模网络架构。此外，优化网络通信时间、提升网络效率也是当前面临的关键挑战。
- 基础设施挑战：智算基础设施的大规模建设带来了对高压直流供电技术、高效液冷散热技术等的刚性需求。
- 运维挑战：大模型训练的高并行和网络化计算特征，对系统的可用性和运维便捷性提出了更高要求。这需要具备快速自动定位故障的能力以及高效的自动断点续训能力[10]。

2　系统设计

在设计智算中心万卡集群的系统时，本文采用了高度优化设计与灵活的网络架构，通过层次化、模块化的策略实现了资源的高效整合与灵活扩展，通过细致规划各个关键层级部署的产品与集成方案，确保了整个集群系统的安全性、可扩展性及高效运行能力。智算中心万卡集群系统整体架构如图1所示，整体架构被划分为3个核心区域：出口区、互联区、业务区。每个区域都承担着特定的功能，共同支撑起一个复杂、灵活的云化智算中心系统。

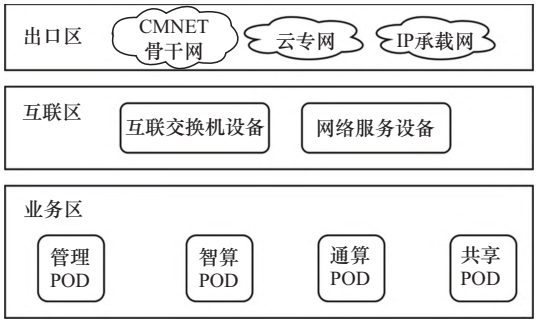


图1　智算中心万卡集群系统整体架构

出口区作为内外部交互的门户，通过高性能路由器和交换机设备实现与中国移动互联网（China Mobile Network，CMNET）骨干网、云专网和互联网协议（Internet protocol，IP）承载网的互联互通。为保障智算中心内部安全，出口区部署了深度包检测（deep packet inspection，DPI）、抗分布式拒绝服务（distributed denial of service，DDoS）系统及防火墙等安全设备，以确保内外网通信的安全性与高效性。

互联区通过部署互联交换机设备，实现了数据中心内部与外部、业务区各交付点（point of delivery，POD）的高速互联互通；通过部署内网防火墙实现资源池内的流量隔离与安全控制，增强内部网络的逻辑分隔与安全性；通过部署软件定义网络（software defined network，SDN）网关设备，如专线网关（transit gateway，TGW）、公网网关（Internet gateway，IGW）、服务网关（customer gateway，CGW）和互联网关（peer gateway，PGW），为上层应用提供公网/专线访问、虚拟私有云（virtual private cloud，VPC）互通、云服务访问等网络服务。

业务区划分为管理POD、智算POD、通算POD和共享POD，以部署不同类型的资源。面对人工智能大模型训练及多样化推理的业务场景，新构建的智算POD网络体系划分为四大关键平面：业务/存储网络平面、管理网络平面、数据网络平面，以及专门针对高性能模型参数同步需求的参数网络平面，智算POD组网拓扑如图2所示。各网络平面均采用叶脊（spine-leaf）网络架构，以提升数据传输效率、降低时延。

2.1 业务面与管理面

业务网络平面主要接入AI服务器、通用服务器和存储混闪服务器。服务器侧配置两个25吉比特以太网（gigabit ethernet，GE）上行端口。组网方式采用两层组网模式，即接入层与汇聚层，网络收敛比为2.5:1。万卡集群接入层交换机采用48×25GE+8×100GE型号。存储混闪服务器共计2×721个25GE端口，共计配置38台Leaf交换机。裸金属网关（bare metal gateway，BMGW）以1:30的比例纳管AI服务器，2 304台AI服务器共计配置154台Leaf交换机。汇聚层部署4台16槽位高性能框式交换机，汇聚层与接入层交换机之间实现交叉全互联。汇聚交换机采取单机部署以简化管理，而接入交换机则成对部署，增加网络冗余度，提高系统稳定性。管理控制方面，部署移动云自主研发的SDN管理系统，配套白盒交换机实现IGW、TGW、CGW、PGW等网关角色，

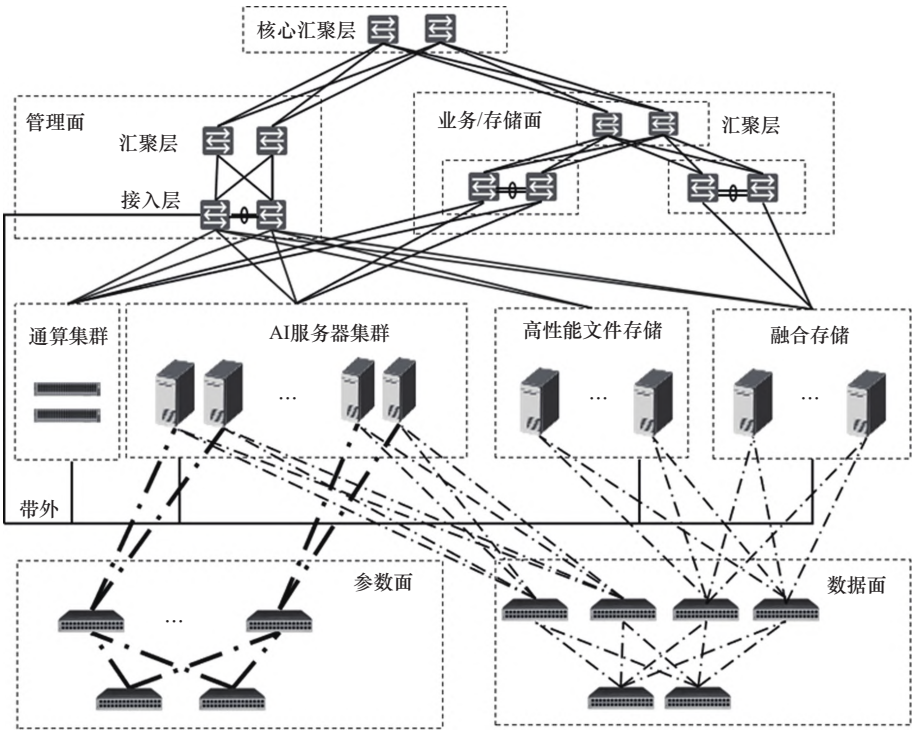


图2 智算POD组网拓扑

承载公网访问、专线接入、存储访问、VPC 互访等业务流量。

管理网络平面主要接入各类型服务器、网络设备和安全设备，采用核心−汇聚−接入的 3 层组网模式，接入层根据业务需求灵活部署 25GE 接入交换机、10GE 接入交换机和 GE 接入交换机等设备。因管理网络方案较为成熟，本文对此不再赘述。

2.2　参数面

参数网络平面接入 2 304 台 AI 服务器，主要用于大模型训练过程中的同步模型梯度和参数，属于带宽敏感性业务流量。接入层的 Leaf 交换机采用 32×400GE 设备，核心层的 Spine 交换机采用高性能 576×400GE 框式设备。Leaf 交换机和 Spine 交换机独立部署，Leaf 设备的上下行收敛比为 1∶1，每台 Leaf 交换机通过 1×400GE 上联至 1 台 Spine 交换机。每台 AI 服务器上行配置为 8×200GE，并采用 Y 型线缆连接，2 个网口共用 1 个 Leaf 交换机的 400GE 端口，即 4 个 400GE 端口连接至 1 台训练服务器。每 4 台 AI 服务器为一组，连接到 1 台 Leaf 交换机，因此，2 304 台服务器需要 576 台 Leaf 设备。576 台 Leaf 设备上行共需要 9 216 个 400GE 端口，每台 576×400GE 则需要配置 16 台 Spine 设备。

2.3　数据面

数据网络平面接入了 AI 服务器和存储集群，主要用于大模型训练数据加载、模型断点保存及恢复、模型保存等场景。接入层的 Leaf 交换机采用 128×100GE 设备。核心层的 Spine 交换机采用高性能 576×100GE 框式设备。为提升集群算力利用率，确保训练计算节点的训练效率，本文采用接入交换机上下行带宽收敛比为 1∶1 的设计，2 304 台 AI 服务器共需要接入 72 台 Leaf 交换机，存储集群共需要上行 3 926×100GE 的带宽，因此，需要 54 台 Leaf 交换机。

2.4　租户隔离

面对云计算场景下智算中心多租户的刚性需求，业务网络通常采用虚拟扩展局域网（virtual extensible local area network，VxLAN）技术来实现 VPC 粒度的网络隔离，该技术已经相对成熟。考虑参数网络的性能因素，本文采用了虚拟局域网（virtual local area network，VLAN）和访问控制列表（access control list，ACL）的方式实现参数网络 VPC 粒度的网络隔离。租户在开通 AI 裸金属服务器时，移动云的 SDN 控制器根据租户业务网络所属的 VPC 来配置参数网络的 VLAN 和 IP 地址。在给网络设备下发 VLAN 和 ACL 配置的同时，SDN 控制器通过裸金属服务将参数网络的 IP 地址配置到 AI 服务器的参数网卡中。在此过程中，SDN 控制器起到了关键的管理与桥梁作用，实现了业务网络与参数网络的统一管理，并确保了这两个网络平面的隔离能力相匹配。

3　系统验证

在分布式深度学习模型训练过程中，通常使用深度学习框架（如 TensorFlow 和 PyTorch）。深度学习框架根据训练任务生成计算任务和集合通信任务，并通过调用集合通信库来实现集合通信原语，以在不同 GPU 之间传输激活函数或同步梯度[11]。All-Reduce 操作是一项常用的集合通信原语，如在 Megatron-LM 框架中，数据并行和一些模型并行策略会使用 All-Reduce 操作[12]。在分布式并行策略中，最直接且应用最广泛的并行策略是数据并行。以数据并行训练为例，All-Reduce 操作用于每次迭代的梯度汇总和同步，以确保所有设备上的模型参数保持一致，因此，数据并行性会导致大量的 All-Reduce 通信[11]。具体来说，Reduce 操作负责将所有 GPU 上的梯度相加求均值，而 Broadcast 操作则将计算结果分发给所有 GPU，以实现参数的同步。All-Reduce 规约操作如图 3 所示，以一个包含 k 张 GPU 卡的求和 All-Reduce 操作为例，每张 GPU 卡提供包含 N 个值的数组，经过 All-Reduce 规约操作后，每张卡对应数组的数值

变成 out[i]=in0[i]+in1[i]+⋯+in(k−1)[i][13]，其中 i 表示数组中元素编号。考虑All-Reduce集合通信在大模型训练过程中最为常见，本文验证了64节点共512卡集群不同组网、带宽等情况下的All-Reduce带宽情况，采用的All-Reduce集合通信算法为HD（halving-doubling）算法。假设参与通信的节点数为N，在通信量相同的情况下，相对于通信步骤为2×(N−1)次的Ring算法，HD算法的通信步骤只有2×lb$N$次[14]。测试方法为采用集合通信测试工具，分别测试1 GB和10 GB数据量集合通信All-Reduce带宽性能。

3.1 参数面网络设计与验证

为验证不同组网方案和网络带宽对集合通信的影响，本文采用单轨组网200 Gbit/s上行和多轨组网100 Gbit/s上行、200 Gbit/s上行3种方案组建了3个不同的人工智能加速卡集群。为控制变量，本文在3个集群中选取了64台机器共512卡组建同等计算节点数量的集群进行测试。每台服务器配置8个上行端口，并采用集合通信测试工具验证3种组网方式下的All-Reduce带宽性能。不同组网方式下的All-Reduce集合通信带宽性能测试结果见表1，在200 Gbit/s端口速率多轨组网情况下，1 GB和10 GB数据量集合通信All-Reduce带宽性能分别为36.7 GB/s和37.3 GB/s。在200 Gbit/s端口速率单轨组网情况下，1 GB和10 GB数据量集合通信All-Reduce带宽性能分别为36.8 GB/s和37.4 GB/s。在100 Gbit/s端口速率多轨组网情况下，1 GB和10 GB数据量集合通信All-Reduce带宽性能分别为26.9 GB/s和27.3 GB/s。在相同带宽情况下，单轨组网和多轨组网情况的性能差距不足1%。在相同组网方案下，单张人工智能加速卡配置200 Gbit/s上行相比100 Gbit/s上行，性能提升约36%。

此外，本文还测试了在PyTorch框架下，基于alpaca-data-conversation.json数据集使用LLaMA-13B模型进行训练的性能对比，并行策略采用张量并行（tensor parallelism，TP）为8，数据并行（data parallelism，DP）为512，3种组网情况下，集群的有效算力差距不足1%。由此可见，在这种跨机并行策略下，100 Gbit/s组网即可满足需求。

为解决Leaf交换机与Spine交换机互联链路等价多路径路由（equal cost multi path routing，ECMP）负载均衡技术存在的哈希冲突和哈希极化问题，本文采用了一分二线缆将Leaf交换机的400GE端口拆分后接入服务器的两个200GE端口，Leaf交换机与Spine交换机互联采用400GE

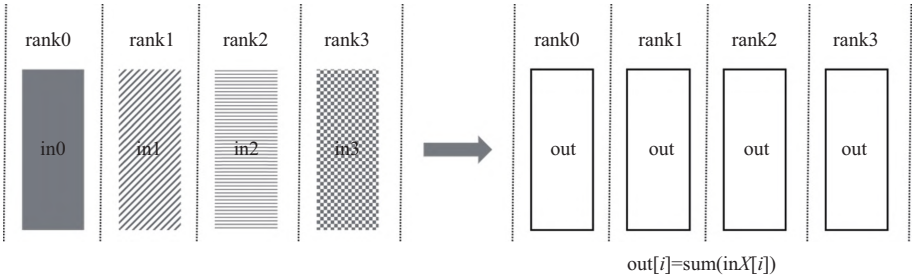


图3 All-Reduce规约操作

表1 不同组网方式下的All-Reduce集合通信带宽性能测试结果

集群配置（64机512卡）	All-Reduce带宽（1 GB数据量）/（GB·s^{-1}）	All-Reduce带宽（10 GB数据量）/（GB·s^{-1}）	有效算力/TFlops
单卡200 Gbit/s上行（多轨组网）	36.7	37.3	122
单卡200 Gbit/s上行（单轨组网）	36.8	37.4	124.4
单卡100 Gbit/s上行（多轨组网）	26.9	27.3	124.2

端口，这样服务器网卡端口带宽和Leaf交换机上行端口带宽比为1:2，即使出现哈希冲突导致两个网卡流量哈希到一个Leaf交换机上行端口时也可以无阻塞转发。此外，本文还采用了基于端口组（rail group）的负载均衡策略。基于端口组的负载均衡策略如图4所示，首先将Leaf交换机与同一服务器互联的端口加入同一个端口组实例中，通过路由转发和接口索引，哈希计算上行链路。将Spine交换机与同一台Leaf交换机的所有互联端口也加入同一个端口组实例中，同样通过路由转发和接口索引，哈希计算下行链路。

本文在64节点512卡集群中验证了开启基于端口组的负载均衡与传统ECMP负载均衡情况下的性能对比测试，ECMP与端口组不同负载均衡策略的带宽利用率如图5所示。测试观测点为Spine设备出入端口的流量统计信息。在采用ECMP负载均衡机制情况下，Spine的16个入端口最大带宽利用率为21.07%，最小带宽利用率为7.02%，16个出端口最大带宽利用率为21.07%，最小带宽利用率为7.02%，可见此时已经出现了哈希冲突。开启基于端口组的负载均衡后，入端口最大带宽利用率为13.84%，最小带宽利用率为13.83%，出端口最大带宽利用率为13.83%，最小带宽利用率为13.83%，可见此方案可有效解决哈希冲突和哈希极化导致的拥塞问题。

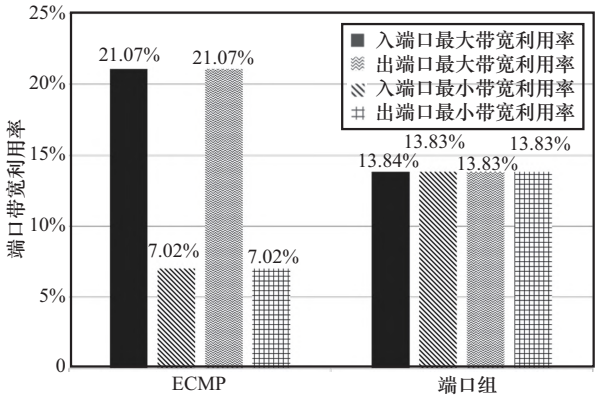


图5　ECMP与端口组不同负载均衡策略的带宽利用率

为充分验证组网规模对集群性能的影响，本文测试了512卡和16 384卡All-Reduce带宽性能，不同规模算力集群All-Reduce带宽性能测试结果见表2。在1 GB数据量情况下，512卡集群平均带宽为36.8 GB/s，16 384卡集群平均带宽为34.5 GB/s。在10 GB数据量情况下，512卡集群平均带宽为37.4 GB/s，16 384卡集群平均带宽为35.0 GB/s。考虑2 048节点16 384卡的集合通信流量存在更多的跨机架、跨核心交换机和跨机房等物理链路因素，整体集合通信性能较64机512卡低6%左右。

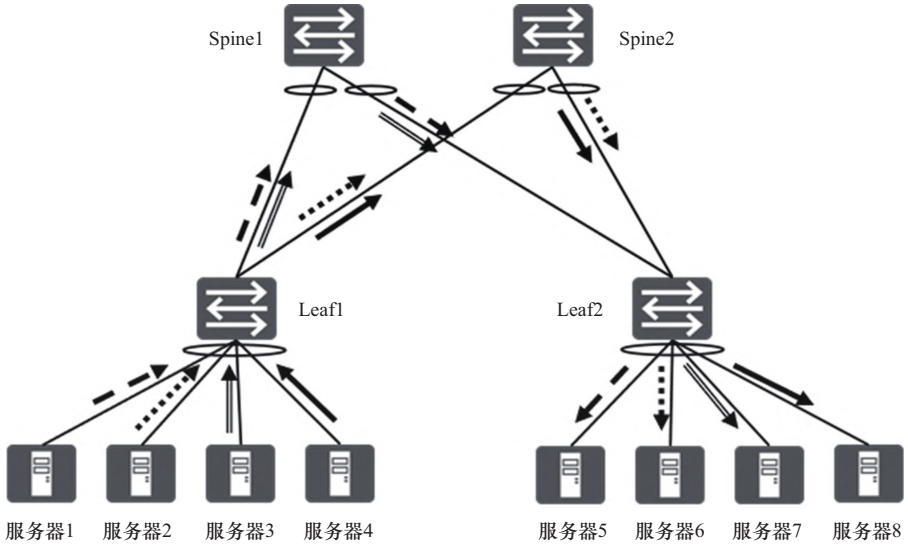


图4　基于端口组的负载均衡策略

表2 不同规模算力集群All-Reduce带宽性能测试结果

集群配置	All-Reduce带宽（1 GB数据量）/（GB·s^{-1}）	All-Reduce带宽（10 GB数据量）/（GB·s^{-1}）
64机512卡	36.8	37.4
2 048机16 384卡	34.5	35.0

3.2 数据面网络设计与验证

在超万卡智算集群中，存储资源的配置主要满足大模型训练过程中的原始数据归集、数据预处理、训练数据加载、检查点（check point，CKP）保存及恢复、模型保存等存储需求，其中数据归集、预处理和模型保存场景下对存储设备容量要求高但对读写性能要求较低，采用传统数据中心网络架构和存储协议即可满足业务需求，本文不再赘述。然而，针对大模型训练过程中训练数据加载、检查点保存及恢复对存储集群的带宽和读写性能都提出了更高的性能要求，因此，本文在系统设计时构建了独立的数据网络平面，数据面高性能存储网络组网架构如图6所示，该网络平面专门用于承载训练数据加载、模型断点保存及恢复的数据流量。

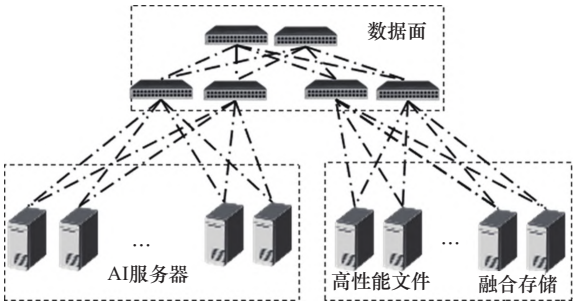


图6 数据面高性能存储网络组网架构

为保证整体集群的性能，在采用高性能存储介质的同时，不同的网络类型和网络协议也会对集群性能产生影响。首先，本文采用支持RDMA的组网方案，使计算机可以远程直接存取其他计算机存储区的数据，而不需要过多的中央处理器（central processing unit，CPU）干预。理论上，通过RDMA互联的计算和存储相较于传统的传输控制协议（transmission control protocol，TCP），实现了百倍的性能提升。此外，为了提供更高的性能，本文采用了一种私有并行客户端存储协议，相对于业内主流采用的网络文件系统（network file system，NFS）协议，该协议具有显著优势。客户端可以直接通过数据面网络将数据写入多个存储节点，不需要存储后端网络做东西向流量转发，从而缩短了数据IO路径长度，性能可以进一步提升50%～100%，不同组网方案和存储协议性能对比见表3。

表3 不同组网方案和存储协议性能对比

网络类型	协议类型	高性能文件带宽读写比6:4/（GB·s^{-1}）
TCP	NFS协议	80
	并行客户端存储协议	125
RDMA	NFS协议	120
	并行客户端存储协议	250

本文使用1 750亿个参数（本文用175B表示）的大型预训练模型（模型大小为3.22 TB）的训练任务，针对训练数据加载和检查点保存及恢复这两个关键场景，在同一套存储设备集群上，通过采用不同的网络类型、协议类型，对比验证存储集群性能。175B模型训练场景下集群性能对比见表4，由表4可知，采用并行客户端存储协议和RDMA网络情况下，相较于使用NFS和TCP的组网方案，每秒的输入输出量（input/output per second，IOPS）读性能提升10.5倍，亿级数据集的加载时长缩短为1/10；模型CKP保存时带宽提高了1.4倍、CKP写入时长缩短为1/2，模型CKP恢复时长缩短为1/3。

4 集群性能验证与运维监控

4.1 集群性能验证

为验证不同参数配置对集群性能指标的影响，本文基于MindSpore框架进行LLaMA2-175B模型训练测试，通过调整不同张量并行（tensor

表4　175B 模型训练场景下集群性能对比

组网	训练数据加载		模型CKP保存		模型CKP恢复	
	IOPS读性能/万	数据加载时长/s	带宽/（GB·s^{-1}）	CKP写入时长/s	带宽/（GB·s^{-1}）	CKP恢复时长/s
NFS+100 Gbit/s TCP	65	154	60	55	85	39
NFS+100 Gbit/s ROMA	85	118	70	47	115	29
并行客户端存储协议+100 Gbit/s TCP	370	27	75	44	130	25
并行客户端存储协议+100 Gbit/s ROMA	750	13	145	23	260	13

parallelism，TP）、流水线并行（pipeline parallelism，PP）、数据并行（data parallelism，DP）、全局批大小（global batch size，GBS）、微批大小（micro batch size，MBS）以及节点数量等参数，测试集群性能指标，LLaMA2-175B 模型训练性能对比见表5。

表5　LLaMA2-175B 模型训练性能对比

参数配置						性能指标/TFlops
TP	PP	DP	GBS	MBS	节点数量	
8	1	128	2 048	2	128	117.96
8	1	32	2 048	2	32	126.51
8	2	16	2 048	2	32	122.35
8	4	8	2 048	2	32	113.25
8	1	64	2 048	2	64	122.61
8	2	32	2 048	2	64	122.19
8	4	16	2 048	2	64	117.41
8	8	8	2 048	2	64	108.34
8	1	64	4 096	2	64	130.72
8	2	32	4 096	2	64	126.98
8	4	16	4 096	2	64	119.76
8	8	8	4 096	2	64	111.44
8	1	64	4 096	2	64	130.72
8	1	64	6 400	2	64	129.58
4	2	64	6 400	2	64	133.12

实验数据显示，在相同计算节点数量情况下，当TP不变时，DP越大，集群平均算力越高，因此，当算力和显存足够大时，应优先选择扩展数据并行模式，减少PP并行。在相同节点数量情况下，当TP、PP、DP策略一致时，GBS越大，集群平均算力越高。

4.2　任务调度及运维监控

整个智算集群部件数量庞大，共包含5 400余台服务器、7.7万条线缆、20万个端口。智算中心承载大模型训练业务呈现的全集群高并行和网络化的计算特征，任一部件故障都会导致训练任务中断，整个集群的高可用和易运维成为智算中心万卡集群可用性的关键指标。因此，本文构建了系统性的任务调度系统和运维监控体系。

在集群性能测试验证过程中，将同一任务随机调度到不同的接入交换机情况下，会小概率出现All-Reduce带宽由35 GB/s下降到28 GB/s的情况，经过排查分析发现，在流量通过Spine交换机转发时，出现了负载不均的情况。为解决该问题，本文增强了智算平台的资源调度引擎，使其支持基于接入交换机和网络拓扑的调度策略，将同一并行任务调度到同一接入交换机设备下。此外，本文增强了设备纳管与故障监控组件，实时监控运行节点状态，预留一定数量的备机，保证故障节点隔离后的断点续训任务被重调度到健康节点，实现断点续训。

本文自主开发了一套实现智算集群设备级别管控的智算管控平台，聚焦共有云多租户智算训练任务的巡检、调优等高阶运维能力，支持不同厂商的计算、存储和网络设备的智算资源的统一管理，支持端网一体化的性能监控和调优能力，实现了芯片健康状态、温度、人工智能加速芯片核心与高带宽显存（high bandwidth memory，

HBM）利用率等关键指标的实时监控。智算管控平台可提供从训前健康检查、训中任务监控，再到训后检查的全栈智能优化服务。本文还建立了故障告警知识库，实现了故障快速定位和自愈。智算管控平台提供了开放的北向接口供智算平台调用，辅助断点续训能力，提升运行效率，确保任务的稳定运行。

5 结束语

本文深入探讨了构建云化的智算中心万卡集群系统的创新与实践，旨在应对人工智能大模型参数量的爆发式增长对智能算力的需求。在对国内外智算中心的发展趋势、建设现状以及面临的挑战进行分析的基础上，本文提出了一种基于国内NPU算力芯片构建的云化智算中心万卡集群系统，并结合项目工程实践对系统中重点的参数网络和存储网络做了系统性的设计及验证。验证结果表明，经过软硬件协同优化，国产化的NPU万卡集群不仅能够满足千亿参数大模型训练的需求，未来更可以支撑万亿参数大模型训练任务，为未来智算中心的发展提供了有力的技术和实践经验支撑。随着团队的调优能力与任务调度系统、运维监控系统功能的不断进步和迭代优化，智算中心万卡集群算力将发挥更出色的性能，未来还将构建更大规模的智算集群，以承载十万亿甚至百万亿参数的模型训练需求。

参考文献：

[1] IDC、浪潮信息、清华全球产业院. 2022—2023全球计算力指数评估报告[R]. 2023.
IDC, Inspur Information, Tsinghua Institute of Global Industry. 2022-2023 global computing index evaluation report[R]. 2023.

[2] 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.

[3] 中国信息通信研究院. 中国算力发展指数白皮书（2022年）[R]. 2022.
China Academy of Information and Communications Technology. White paper on China's computing power development index (2022)[R]. 2022.

[4] AN W, BI X, CHEN G T, et al. Fire-Flyer AI-HPC: a cost-effective software-hardware co-design for deep learning[J]. arXiv preprint 2024: 2408.14158v1.

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

[6] 腾讯云开发者社区. 大模型在机器学习领域的运用及其演变: 从深度学习的崛起至生成式人工智能的飞跃[EB]. 2024.
Tencent cloud developer community. The application and evolution of large models in machine learning: from the emergence of deep learning of artificial intelligence to generate the type leap[EB]. 2024.

[7] 阿里云开发者社区. 10万亿！达摩院发布全球最大AI预训练模型M6[EB]. 2021.
Alibaba Cloud Developer Community. 10 trillion! damo academy releases the world's largest AI pre-trained model M6 [EB]. 2021.

[8] 通信产业网. 万卡集群: 为什么？是什么？怎么建？[EB]. 2024.
Communications Industry Network. Wan Ka Cluster: Why? What? How to Build? [EB]. 2024.

[9] 鲍中帅. 万卡级超大规模智算集群网络运维挑战及实战[EB]. 2024.
BAO Z S. The operation and maintenance challenge and actual combat of super scale intelligent computing cluster network [EB]. 2024.

[10] 中国移动. 面向超万卡集群的新型智算技术白皮书（2024年）[R]. 2024.
China Mobile. For a new type of wisdom is above all card cluster technology, the white paper (2024) [R]. 2024.

[11] WEI Y Z, HU T S, LIANG C, et al. Communication optimization for distributed training: architecture, advances, and opportunities[J]. arXiv preprint，2024: 2403.07585.

[12] SHOEYBI M, PATWARY M, PURI R, et al. Megatron-LM: training multi-billion parameter language models using model parallelism[J]. arXiv preprint，2019: 1909.08053.

[13] NVIDIA Deep Learning NCCL Documentation. Collective op-

erations[EB]. 2024.

[14] DONG J B, WANG S C, FENG F, et al. ACCL: architecting highly scalable distributed training systems with highly efficient collective communication library[J]. IEEE Micro, 2021, 41(5): 85-92.

[作者简介]

丁宏庆（1972– ），男，中国移动通信集团有限公司计划建设部副总经理，主要研究方向为算力网络、云计算及人工智能技术等。

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。

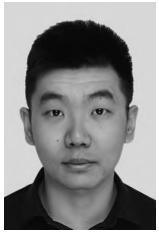
牛红韦华（1989– ），女，中移（苏州）软件技术有限公司计划建设部总经理助理，主要研究方向为大规模智算中心、公有云资源池架构、云管理平台等。

李志勇（1986– ），男，中国移动通信集团浙江有限公司高级工程师，主要研究方向为云计算和人工智能技术等。

周丹媛（1990– ），女，中国移动通信集团有限公司高级项目经理，主要研究方向为云计算和人工智能技术等。

丁国强（1986– ），男，中国移动通信集团设计院有限公司高级研究专员，主要研究方向为云计算和人工智能技术等。

李攀攀（1990– ），男，中移（苏州）软件技术有限公司软件架构师，主要研究方向云计算、人工智能技术等。



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

张久仙（1988– ），男，中移（苏州）软件技术有限公司解决方案架构师，主要研究方向云计算、高性能网络、人工智能技术等。