中国移动 China Mobile

中移智库

面向超万卡集群的新型智算技术白皮书

（2024年）

中国移动通信集团有限公司

编制单位：中移智库

前 言

自 ChatGPT 发布以来，科技界掀起了一场大模型的竞争热潮。数据成为新生产要素，算力成为新基础能源，大模型则成为新生产工具，各行各业从“+AI”向“AI+”的转变已势不可挡。随着模型参数量从千亿迈向万亿，模型能力更加泛化，大模型对底层算力的诉求进一步升级，超万卡集群成为这一轮大模型基建军备竞赛的标配。

超万卡集群将有助于压缩大模型训练时间，实现模型能力的快速迭代，并及时对市场趋势作出应对。然而，如何在超万卡集群中实现高效的训练，并长期保持训练过程的稳定性，是将大模型训练扩展到数万张 GPU 卡上所要面临的双重挑战。超万卡集群运行过程中涉及到集群有效算力发挥、超大规模互联网络稳定性保障、故障的快速排查和修复等关键问题，目前都是业内关注的焦点。

中国移动全面拥抱“AI+”时代，提出超万卡集群的核心设计原则，并在计算、存储、网络、平台及机房配套等多个领域提出关键问题和解决方案。中国移动希望与行业一起应对超万卡集群所带来的前所未有的挑战，共同助推国内智算基础设施迈向新的台阶。

目 录

第一章：超万卡集群背景与趋势

1.1 大模型驱动智能算力爆发式增长

自 ChatGPT 面世以来，大模型步入了迅猛发展期，模型层出不穷，爆点频出，Scaling Law[1]不断得到验证，高速发展的人工智能对数字经济产生了巨大赋能作用。大模型所使用的数据量和参数规模呈现“指数级”增长，2018 年 BERT 模型参数量仅有 1.1 亿，到 2021 年 GPT-3 达到了 1750 亿。随着 Mixture of Experts（MoE）[2]等先进模型结构的出现，模型参数迈入万亿规模。预计在未来的 2-3 年，随着 AI 技术的进步和算力提升，Scaling Law 还将延续，助推模型参数向十万亿进军。

大模型能力的不断跃迁，使得超长序列应用、文生视频、文生音频等基于多模态的应用层出不穷，大模型在各个领域均展现出了强大的智能化能力，“AI+”对生产生活带来了巨大影响。ChatGLM、LLaMA[3]、Gemini 等大模型的发布更加坚定了科技界持续摸高大模型能力天花板的决心；文生视频多模态大模型 Sora 的问世更加引爆了行业热点，促使业界在大模型的技术、规模和应用上不断挖掘，以期能创造下一轮爆点。

AI 技术的发展带动产业大规模升级的同时，也带来了对巨量算力和能源的需求。据公开信息报道，GPT-3 训练所消耗的电力，相当于美国约 121 个家庭一整年的用电量。GPT-4 拥有 16 个专家模型共 1.8 万亿参数，一次训练需要在大约 25000 个 A100 上训练 90 到 100 天。大模型对底层算力、空间、水电能源产生极大消耗，对新一代智算设施的设计要求也日益严苛。更高密度的算存硬件、高性能无阻塞的网络连接以及更高并行度的通信和计算范式成为新一代智算中心的设计目标，新型智算中心（NICC，New Intelligent Computing Center）[4]相关技术将继续被推向新的高度。

1.2 超万卡集群的建设正在提速

人工智能新纪元，算力为企业科技创新和转型提供有力支撑。在全球化的科技竞争格局中，领先的科技公司正积极部署千卡乃至超万卡规模的计算集群，既是展现其在人工智能、数据分析、大模型研发等前沿领域的技术实力，也向外界展示了公司对

未来科技趋势的深远布局。

在国际舞台上，诸如 Google、Meta、Microsoft 等科技巨头，正利用超万卡集群推动其在基座大模型、智能算法研发及生态服务等方面的技术创新。如 Google 推出超级计算机 A3 Virtual Machines，拥有 26000 块 Nvidia H100 GPU，同时基于自研芯片搭建 TPUv5p 8960 卡集群。Meta 在 2022 年推出了一个拥有 16,000 块 Nvidia A100 的 AI 研究超级集群 AI Research Super Cluster, 2024 年初又公布 2 个 24576 块 Nvidia H100 集群，用于支持下一代生成式 AI 模型的训练。这些企业通过成千上万台服务器组成的集群计算优势，不断优化服务架构，提升用户体验，加速新技术的市场转化与应用。

在国内，通信运营商、头部互联网、大型 AI 研发企业、AI 初创企业等均在超万卡集群的建设和使用过程中不断推动技术革新。

（一） **通信运营商作为国家算力基础设施建设的中坚力量，利用其庞大的机房资源和配套设施优势，正加速推进超万卡集群智算中心的建设**。这一举措不仅为运营商自身的大模型研发提供强大的计算支持，同时也为政府、高校和企业客户带来了前所未有的高质量智算服务。随着智算中心建设的不断深入，运营商站在连接技术创新与行业应用的关键位置，其在推动社会数字化转型和智能化升级中的引领作用日益凸显。

（二） **头部互联网企业作为技术创新的先锋，通过建设超万卡集群来加速其在云计算、大数据分析和大模型研发等领域的突破**。字节跳动、阿里巴巴、百度为代表的互联网公司在积极推进超万卡集群的建设。其中，字节跳动搭建了一个 12288 卡 Ampere 架构训练集群，研发 MegaScale 生产系统用于训练大语言模型[5]。通过集群的强大计算力，这些头部互联网公司不仅加速了自身业务的数字化转型，也为国内科技产业的发展贡献了创新动力。

（三） **大型 AI 研发企业出于对大规模模型训练和复杂算法计算的迫切需求，正在积极投建超万卡集群**。这些公司作为人工智能领域的先行者，正积极投建超万卡集群以满足其大模型的计算需求。如科大讯飞，2023 年建设成首个支持大模型训练的超万卡集群算力平台“飞星一号”。此类集群的建立，不仅为这

些企业在 AI 领域的深入研究提供了必须的算力支撑，也为他们在智算服务的商业应用中赢得了先机。

（四） **AI 初创企业则更倾向于采取灵活的租用模式，利用已有的超万卡集群来支持其创新项目**。这些企业为了能够实现应用和投入平衡，大多对基础设施采取灵活的租用模式，利用超万卡集群的强大计算能力来支持其创新项目。这种模式降低了初创企业的初始投资门槛，使他们能够快速获得高性能的计算资源，加速产品的研发和迭代。

整体而言，无论是通信运营商、头部互联网企业、大型 AI 研发企业还是 AI 初创企业，都在通过自建或使用超万卡集群加速其在人工智能领域的技术突破和产业创新。随着超万卡集群建设的不断深入，我们预见这一趋势将为整个智算产业的发展带来深远影响。

第二章：超万卡集群面临的挑战

当前，超万卡集群的建设仍处于起步阶段，主要依赖英伟达 GPU 及配套设备实现。英伟达作为全球领先的 GPU 供应商，其产品在大模型训练上有较大优势。得益于政策加持和应用驱动，国产 AI 芯片在这两年取得长足进步，但在整体性能和生态构建方面仍存在一定差距。构建一个基于国产生态体系、技术领先的超万卡集群仍面临诸多挑战。

2.1 极致算力使用效率的挑战

大量实践表明，针对大模型分布式训练场景，集群规模的线性提升无法直接带来集群有效算力的线性提升，卡间和节点间的互联网络、软件和硬件的适配调优是追求集群极致有效算力的关键挑战。我们把集群有效算力分解为“**GPU 利用率**”和“**集群线性加速比**”两个重要指标，其中“GPU 利用率”受限于芯片架构和制程、内存和 I/O 访问瓶颈、卡间互联带宽和拓扑、芯片功耗等因素，“集群线性加速比”则取决于节点间的通信能力、并行训练框架、资源调度等因素，两者的最大化发挥将最终表现为模型训练效率提升和成本降低。在超万卡集群中，需要运用系统工程方法，通过对超万卡集群网络的精细化设计、软硬件全栈整合优化，综合提升集群算力使用效率。

2.2 海量数据处理的挑战

千亿模型的训练需要对 PB 量级的数据集使用多种协议进行处理，未来万亿模型的训练对 checkpoint 的读写吞吐性能更是要求高达 10TB/s，现有智算存储系统在协议处理、数据管理、吞吐性能等方面面临诸多挑战:

- **协议处理层面**：传统智算存储系统按照块、文件、对象等不同协议建设分立存储池，多套不同协议存储系统之间需要来回拷贝数据，影响数据处理效率，浪费存储空间，增加运维难度；
- **吞吐性能层面**：传统智算的分布式文件存储仅支持百节点级别扩展，节点规模小，难以提供超万卡集群所需的 10TB/s 以上的数据吞吐性能；

- **数据管理层面**：传统智算的数据存储需人工干预，进行冷热分类，并在高性能和普通性能存储集群之间迁移。跨系统的数据管理和迁移降低了大模型下海量数据处理效率，还会额外占用网络带宽和计算节点资源。

因此，超万卡集群的存储系统需要通过协议融合、自动分级等一系列技术手段，提供高效的数据共享和处理能力，满足大模型训练的需求。

2.3 超大规模互联的挑战

模型规模扩大到万亿量级，数据的吞吐量和计算量已远远超过目前最强单机单卡能力，多机多卡互联和并行训练策略成为必须。以在超万卡集群部署 1.8 万亿 GPT-4 为例，在大模型训练过程中，每轮迭代计算都涉及前反向传播算法的计算和通信，这对超万卡集群的 Scale Out 和 Scale UP 网络提出极大挑战。

- **在 Scale Out 互联层面，**网络承载数据并行（Data Parallel，DP）和流水线并行（Pipeline Parallel，PP）流量，参数面网络带宽需达到 200Gbps 至 400Gbps，数据面网络需要配备 100Gbps 带宽，保证数据读取不成为训练瓶颈。此外，参数面网络还需要应对因多租户多任务并行训练通信特征不规整、上下行 ECMP（Equal Cost Multi Path）选路不均衡而引发的高速大象流的交换冲突和拥塞。
- **在 Scale up 互联层面，**由于 MoE 专家并行和张量并行（Tensor Parallel，TP）的通信无法被计算掩盖，不仅要求卡间互联带宽达到几百甚至上千 GB 的量级，而且应突破当前单机 8 卡的限制，以支持更大参数量的模型训练。此外，Scale up 互联还需要保持高频度、低时延、无阻塞的通信模式。

2.4 集群高可用和易运维挑战

超万卡集群承载万亿模型训练意味着千万器件的满负荷高速运转，任一部件不可恢复的失效都可能导致训练中断，带来超万卡集群高可用和易运维的关键挑战：

- **千万器件维护管理难度大：**超万卡集群由数千台智算服务器+数千台交换机+数千台存储设备以及数万根光纤/数万颗光模块构成，训练任务涉及千万颗元器

件满负荷高速运转，基于固有的元器件硬件失效率和海量的器件规模带来硬件故障频发，涉及到的软硬件故障模式繁杂，故障管理挑战巨大；

- **复杂系统故障定位难度大：**万亿模型训练的过程是各个软硬组件精密配合的过程，一旦发生问题定界定位复杂。业界典型硬件故障定位需 1~2 天，复杂应用类故障定位可能长达数十天。快速自动定界定位能力需要结合实际运维经验进行系统性积累和针对性持续改进。
- **高负荷运行故障概率高：**万亿大模型训练至 TTA（Time To Accuracy）一般需要一百天 7×24 小时满负荷运行。而硬件 MTBF（Mean Time Between Failure）伴随集群规模的增长越来越短，万亿大模型训练作业中断频发，业界超万卡集群持续稳定运行仅数天，断点续训恢复缓慢，直接影响模型训练效率。超万卡集群急需支持更有效、更快速、影响更小的自动断点续训功能。

2.5 高能耗高密度机房设计的挑战

超万卡集群对机房配套设施的需求相对于传统 IDC 云数据中心发生重大变化，对供电、承重、机房洁净度和走线架设计等有极高要求：

- **在供电方面**，当芯片 TDP 设计功率上升至 400~700W，单柜功率从原先的 7~8KW 上升至 40KW 甚至 60KW，集群整体功耗将攀升至数十~上百 MW，机房需要进行功率提升改造，并配合进行散热能力提升改造；
- **在承重方面**，由于集群规模翻番增长，为了保障单位空间的算力密度，需要引入液冷方案，确保智算芯片的高速运行，单机柜重量达 1-2 吨，对机房承重提出高标准要求；
- **在机房洁净度方面**，由于超万卡集群参数面网络使用大量 100G、200G 甚至 400G 的高速光模块，高速光模块本身是一个集成度极高的器件，裸露的光纤通道和内部器件都比较脆弱，要尽可能避免落入灰尘，降低故障率。因此机房需要综合考量制冷和通风方案，在设备侧保持较高的洁净度标准，确保后期集群的稳定运行。
- **在线缆布放方面**，由于超万卡集群的算力密度更高、功耗密度更高，线缆的布

放量也随之增大。以一个 1.8 万卡的智算集群为例，需要布放 10 万量级的线缆，这将对走线架的宽度和承重带来新的挑战。

可见，超万卡集群提出了对高压直流供电技术、高效液冷散热技术、超大规模网络工程便捷落地的刚性需求。这就要求机房配套设施在建设之初，提前对机房供电、制冷、承重等进行配套设计，以更好的支撑超万卡集群快速建设、便捷部署和长期稳定运行。

第三章：超万卡集群的核心设计原则和总体架构

3.1 超万卡集群的核心设计原则

在大算力结合大数据生成大模型的发展路径下，超万卡集群的搭建不是简简单单的算力堆叠，要让数万张 GPU 卡像一台“超级计算机”一样高效运转，超万卡集群的总体设计应遵循以下五大原则：

- **坚持打造极致集群算力：**基于 Scale-up 互联打造单节点算力峰值，基于 Scale-out 互联将单集群规模推高至万卡以上，两者叠加构建超万卡集群的大算力基座；
- **坚持构建协同调优系统：**依托超大规模的算力集群，通过 DP/PP/TP/EP 等各种分布式并行训练策略，持续提升有效算力，实现极致的计算通信比，最大化模型开发效率；
- **坚持实现长稳可靠训练：**具备自动检测和修复软硬件故障的能力，面向千万器件满负荷运行系统，持续提升 MTBF 和降低 MTTR 并实现自动断点续训能力，支持千亿稠密、万亿稀疏大模型百天长稳训练，保证系统稳定性和鲁棒性；
- **坚持提供灵活算力供给：**支持集群算力调度，提供灵活弹性的算力供给和隔离手段，实现训练和推理资源的按需调配，保持单集群大作业和多租户多任务并行训练性能持平；
- **坚持推进绿色低碳发展：**持续推进全套液冷解决方案在超万卡集群的应用，追求极致绿色算力能效比（FLOPs/W）和极低液冷 PUE 至 1.10 以下。

3.2 超万卡集群的总体架构设计

超万卡集群的总体架构由四层一域构成（如图 1），四层分别是机房配套、基础设施、智算平台和应用使能，一域是智算运营和运维域。

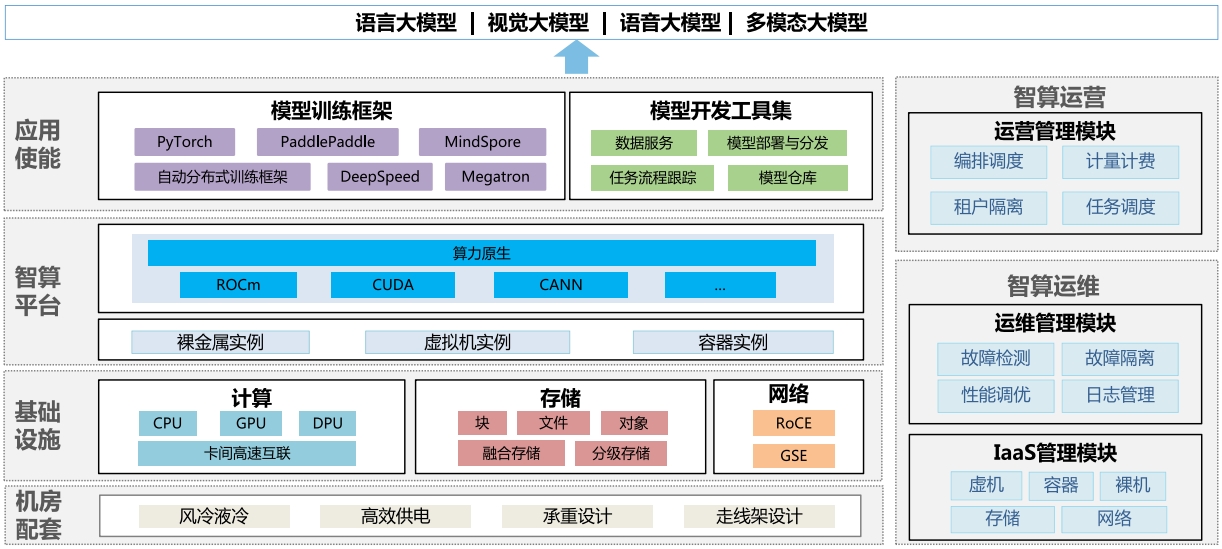

图 1 面向超万卡集群的新型智算总体架构设计

- **机房配套层**：匹配超万卡集群高密集约的建设模式，机房配套设施需重点考虑高效供电、制冷设计、楼板承重和走线架设计等。
- **基础设施层**：算、网、存三大硬件资源有机配合，达成集群算力最优。面向算力，CPU、GPU、DPU 三大芯片协同，最大化发挥集群计算能力；面向网络，参数面、数据面、业务面、管理面独立组网，参数面/数据面采用大带宽 RoCE 交换和二层无阻塞 CLOS 组网满足大象流，支持参数面负载均衡和多租安全隔离；面向存储，引入融合存储和分级存储支持无阻塞数据并发访问。
- **智算平台层**：采用 K8s，对上提供以裸金属和容器为主的集群资源。在对集群资源进行纳管的基础上，进一步实现大规模集群的自动化精准故障管理，以达成高效训练、长稳运行的目标。面向未来，考虑集群中引入异厂家 GPU 芯片，为避免智算碎片化问题，引入算力原生，实现应用跨架构迁移和异构混训等平台能力。
- **应用使能层**：包括模型训练框架和开发工具集两个模块，一方面基于现有开源框架能力，进行分布式训练调优，面向未来开展自动分布式训练框架设计，积累经验，实现对通信和计算重叠的优化、算子融合以及网络性能的高效调优；另一方面，研发沉淀数据服务、模型部署开发等工具集，逐步实现由人工处理到基于工具对外提供自动化模型研发能力的转变。
- **智算运营和运维域**：支持超万卡集群高效集合通信和调度。支持按租户灵活资源发放和任务调度，支持多任务并行训练。

第四章：超万卡集群关键技术

4.1 集群高能效计算技术

随着大模型从千亿参数的自然语言模型向万亿参数的多模态模型升级演进，超万卡集群亟需全面提升底层计算能力。具体而言，包括增强单芯片能力、提升超节点计算能力、基于 DPU（Data Processing Unit）实现多计算能力融合以及追求极致算力能效比。这些系统性的提升将共同支持更大规模的模型训练和推理任务，满足迅速增长的业务需求。

4.1.1 单芯片能力

超万卡集群中，单芯片能力包括单个 GPU 的计算性能和 GPU 显存的访问性能。

在单个 GPU 计算性能方面，首先需要设计先进的 GPU 处理器，在功耗允许条件下，研发单 GPU 更多并行处理核心，努力提高运行频率。其次，通过优化高速缓存设计，减少 GPU 访问内存延迟，进一步提升单 GPU 芯片运行效率。第三，优化浮点数表示格式，探索从 FP16 到 FP8 浮点数的表示格式，通过在芯片中引入新的存储方式和精度，在保持一定精度条件下，大幅提升计算性能。最后，针对特定计算任务，可在 GPU 芯片上集成定制化的硬件加速逻辑单元，这种基于 DSA（Domain Specific Architecture）的并行计算设计，可提升某些特定业务领域的计算速度。

在 GPU 显存访问性能方面，为了将万亿模型的数据布放在数万张 GPU 显存上，要求显存支持高带宽、大容量的能力，确保计算单元能够高效完成访存任务，维持系统的低能耗运行。为便捷访问显存数据，建议 GPU 显存采用基于 2.5D/3D 堆叠的 HBM 技术[6]，减少数据传输距离，降低访存延迟,提升 GPU 计算单元与显存之间的互联效率。

通过这些技术的实施，超万卡集群不仅能够为智算中心提供强大的单卡算力处理能力，还能为未来更大规模的模型训练和推理任务奠定坚实的硬件基础。

4.1.2 超节点计算能力

针对万亿模型的训练与推理任务，特别是在超长序列输入和 MoE 架构的应用背景下，应重点优化巨量参数和庞大数据样本的计算效率，满足由此引发的 All2All 通

信模式下的 GPU 卡间通信需求。为此，建议超万卡集群的改进策略集中在以下几个关键领域：

- 加速推进超越单机 8 卡的超节点形态服务器

为满足万亿或更大参数量模型的部署需求，建议产业界致力于研制突破单机 8 卡限制的超节点形态服务器，通过利用提高 GPU 南向的 Scale up 互联能力，提升张量并行或 MoE 并行对大模型训练任务的收益，实现性能跃升，缩短训练总时长，实现大模型训练整体性能的优化。

- 加快引入面向 Scale up 的 Switch 芯片

建议在节点内集成支持 Scale up 能力的 Switch 芯片，以优化 GPU 南向的互联效率和规模，增强张量并行或 MoE 并行的数据传输能力。如图 2 所示，通过引入节点内的 Switch 芯片，以增强 GPU 卡间的点对点（Point to Point，P2P）带宽，有效提升节点内的网络传输效率，满足大模型日益增长的 GPU 互联和带宽需求，为大规模并行计算任务提供强有力的硬件支持。

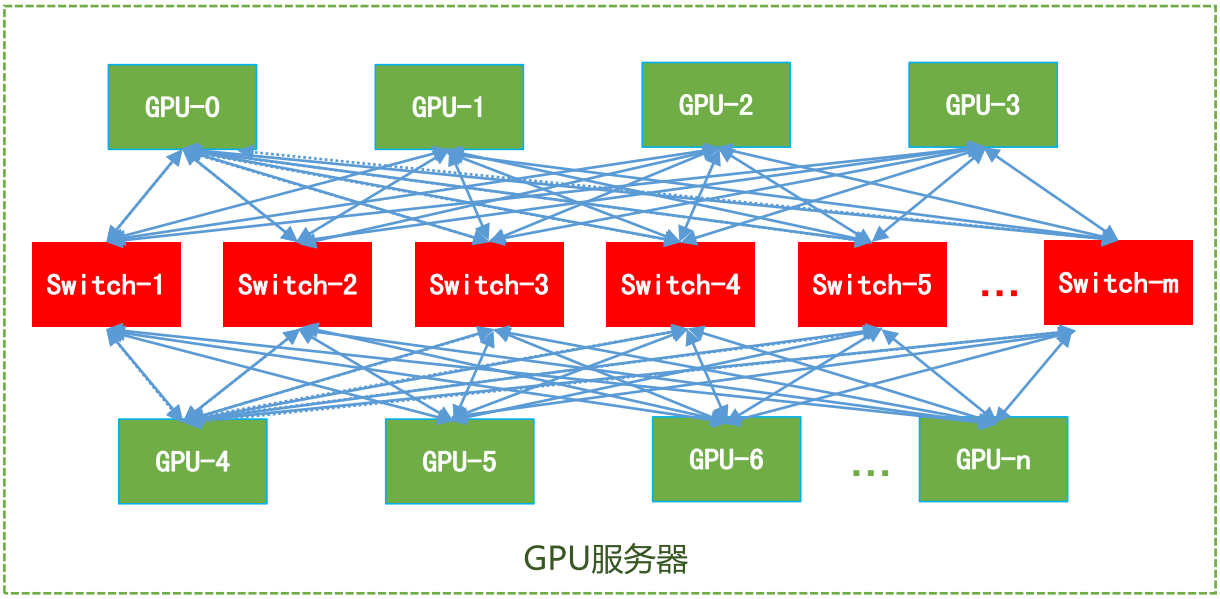

图 2 在服务器内部引入 Switch 芯片示例

- 优化 GPU 卡间互联协议以实现通信效率跃升

建议对 GPU 卡间互联协议进行系统性优化和重构，以提升 All2All 模式下的通

信效率。通过重新设计卡间通信过程中的数据报文格式、引入 CPO（Co-Packaged Optics）/NPO（Near Packaged Optics）、提高和优化 SerDes 传输速率、优化拥塞控制和重传机制以及多异构芯片 C2C（Chip-to-Chip）封装等多种途径，提高超万卡集群的 GPU 卡间互联的网络利用率，减少通信时延，实现带宽能力跃升，从而支持所需的更高频次、更大带宽和更低延迟通信特性。

4.1.3 多计算能力融合

面向超万卡集群，考虑到智算中心内部成倍增长的数据交换需求，通过堆叠 CPU 资源来处理网络数据的做法无疑是低效且昂贵的，对此，智算中心的计算架构需要转变方向，将原本运行在 CPU、GPU 中的数据处理任务卸载至具有层级化可编程、低时延网络、统一管控等特性的 DPU 上执行，在大幅扩展节点间算力连接能力的同时，释放 CPU、GPU 的算力，降低节点间的 CPU、GPU 协作成本，支撑集群发挥更大的效能。

具体地，可以对智算中心进行软硬一体重构，打造计算、存储、网络、安全、管控五大引擎，定义标准化的 DPU 片上驱动内核：

- 计算引擎卸载加速 I/O 设备的数据路径与控制路径，面向节点提供标准化的 virtio-net(Virtual I/O Network)、virtio-blk(Virtiual I/O block)后端接口，屏蔽厂商专用驱动。
- 存储引擎在 DPU 上实现存储后端接口，可基于传统 TCP/IP 网络协议栈或 RDMA(Remote Direct Memory Access)网络功能连接块存储集群、对象存储集群、文件存储集群及文件存储集群，将节点的全类型存储任务卸载至 DPU 中完成。
- 网络引擎将虚拟交换机卸载至 DPU 上，采用标准的流表和卸载接口实现网络流量的卸载，全线速释放硬件性能；同时集成 RDMA 网络功能，降低多机多卡间端到端通信时延，提升多机间端到端通信带宽至 400G 级别，构建节点间数据交换的“高速通道”。
- 安全引擎通过信任根机制以及标准的 IPsec 等加密通讯协议对系统和多租户网络进行安全防护，并基于 DPU 提供有效的卸载方案。
- 管控引擎屏蔽裸金属、虚拟机和容器等算力单元的形态差异，实现 DPU 资源

统一管理和全链路管控运维。

- 以上述五大引擎为蓝图，中国移动于 2020 开始打造具有自主知识产权的磐石 DPU，并于 2021 年正式推出磐石 DPU 版本。经过移动云现网的打磨，中国移动持续升级磐石 DPU 产品能力，并于 2024 年将磐石 DPU 的 FPGA 架构全面升级为 ASIC 架构，旨在围绕磐石 DPU 软硬融合重构算力基础设施，重新定义算力时代云计算技术新标准，构建算力时代新技术曲线。

将以磐石 DPU 为代表的 DPU 芯片融入现有智算中心技术体系，将算力集群由 CPU+GPU 双平台支撑扩展至由 CPU+GPU+DPU 三平台支撑，可以有效联合集群节点间因数据 I/O 瓶颈而产生的算力孤岛，突破现有技术架构下的集群规模极限，使超万卡集群成为可能。

4.1.4 极致算力能效比

在制程工艺相对固定的条件下，芯片的高性能无疑会增加芯片的功耗，从而影响整机的散热。面对高性能计算芯片功率密度急剧上升的现状，需要通过制冷系统和 GPU 芯片两方面进行优化。

在制冷系统方面，当前单机 8 卡 GPU 服务器功耗已经数倍于通用服务器，由于 GPU 的散热量大幅增加，为了增加计算密度，节省空间，超万卡集群建议采用当前较成熟的高密度冷板式液冷机柜，一个液冷机柜可容纳多台液冷 GPU 训练服务器，相比传统风冷机柜大幅提升空间利用率。

在 GPU 芯片方面，为了提升 GPU 单芯片的能效比，应采取多领域的优化策略，实现高性能与低能耗之间的平衡。在芯片工艺领域，建议采用更加先进的半导体制造工艺，如 7nm 或更小的特征尺寸，以此降低晶体管的功耗，同时提升单芯片集成度。此外，应加强超万卡集群内 GPU 架构的创新设计，包括优化片上总线设计、改进流水线结构、优化电压和频率策略以及精确的时钟门控技术，从而在不同工作状态下实现最优的能耗效率。在软件层面，超万卡集群应采用更加精细的监控和分析，实时跟踪 GPU 的运行数据，并不断优化算法和工作负载分配，以实现更加均衡和高效的算力利用。通过上述设计和优化，不仅能提高用户的计算体验，降低成本，也为智算中心可持续发展和绿色环保提出了可行方案。

4.2 高性能融合存储技术

为了实现存储空间高效利用、数据高效流动，并支持智算集群大规模扩展，超万卡集群应采用多协议融合和自动分级存储技术，提升智算数据处理效率，助力超万卡集群支撑千亿乃至万亿大模型训练。

4.2.1 多协议融合

超万卡集群融合存储底座承载 AI 全流程业务数据处理，兼容 AI 全流程工具链所需的 NFS（Network File System）、S3（Sample Storage Service）和并行客户端 POSIX（Portable Operating System Interface）等协议，支持各协议语义无损，达到与原生协议一样的生态兼容性要求，在不同阶段实现数据零拷贝和格式零转换，确保前一阶段的输出可以作为后一阶段的输入，实现 AI 各阶段协同业务的无缝对接，达到“零等待”效果，显著提升大模型训练效率。

4.2.2 集群高吞吐性能

为满足超万卡集群大模型对于存储高吞吐性能需求，基于全局文件系统技术，可支持超 3000 节点扩展规模，为大模型训练提供百 PB 级全闪存储大集群能力，从闪存密度、数据面网络、并行客户端和对等通信机制等多个维度全面提升存储系统性能，实现存储集群 10TB/s 级聚合吞吐带宽、亿级 IOPS，智能算力利用率提升 20%以上，大模型 checkpoint 恢复时长从分钟级提升至秒级，同时对高价值智算存储数据提供强一致性访问和 99.9999% 可靠性能力。

4.2.3 高效分级管理

超万卡集群数据量巨大，其中大部分是温冷数据，统筹考虑性能和成本因素，规划普通性能、高性能两类存储集群。普通性能存储集群使用混闪存储介质，具备低成本和大容量优势，提供温冷数据存储；高性能存储集群使用全闪存储介质，为大模型训练提供数据高吞吐能力，主要用于存放热数据。为智算应用高效管理和访问数据，两类存储集群应该对外呈现统一命名空间，提供基于策略的数据自动分级流动能力，实现冷热数据按照策略自动流动，避免人工频繁介入，提升存储系统整体运行效率。

4.3 大规模机间高可靠网络技术

超万卡集群网络包括参数面网络、数据面网络、业务面网络、管理面网络。业务面网络、管理面网络一般采用传统的 TCP 方式部署，参数面网络用于计算节点之间参数交换，要求具备高带宽无损能力。数据面网络用于计算节点访问存储节点，也有高带宽无损网络的诉求。超万卡集群对参数面网络要求最高，主要体现在四个方面：大规模，零丢包，高吞吐，高可靠。

目前业界成熟的参数面主要包括 IB（InfiniBand）和 RoCE 两种技术。面向未来 AI 大模型演进对网络提出的大规模组网和高性能节点通信需求，业界也在探索基于以太网新一代智算中心网络技术，包括由中国移动主导的全调度以太网（Global Scheduled Ethernet，GSE）方案[6]和 Linux Foundation 成立的超以太网联盟(Ultra Ethernet Consortium,UEC)，两者通过革新以太网现有通信栈，突破传统以太网性能瓶颈，为后续人工智能和高性能计算提供高性能网络。中国移动也将加速推动 GSE 技术方案和产业成熟，提升 AI 网络性能，充分释放 GPU 算力，助力 AI 产业发展。

4.3.1 大规模组网

根据不同的 AI 服务器规模，参数面网络推荐采用 Spine-Leaf 两层组网或胖树（Fat-Tree）组网。

Spine-Leaf 两层组网如图 3 所示。每 8 台 Leaf 交换机和下挂的 AI 服务器做为一个 group，以 group 为单位进行扩展。在 group 内部，推荐采用多轨方案将 AI 服务器连接至 Leaf 交换机，即所有 AI 服务器的 1 号网口都上连至 Leaf1，所有 2 号网口上连至 Leaf2，依此类推，所有 8 号网口上连至 Leaf8。Spine 交换机和 Leaf 交换机之间采用 Fullmesh 全连接。Leaf 交换机上下行收敛比为 1:1。

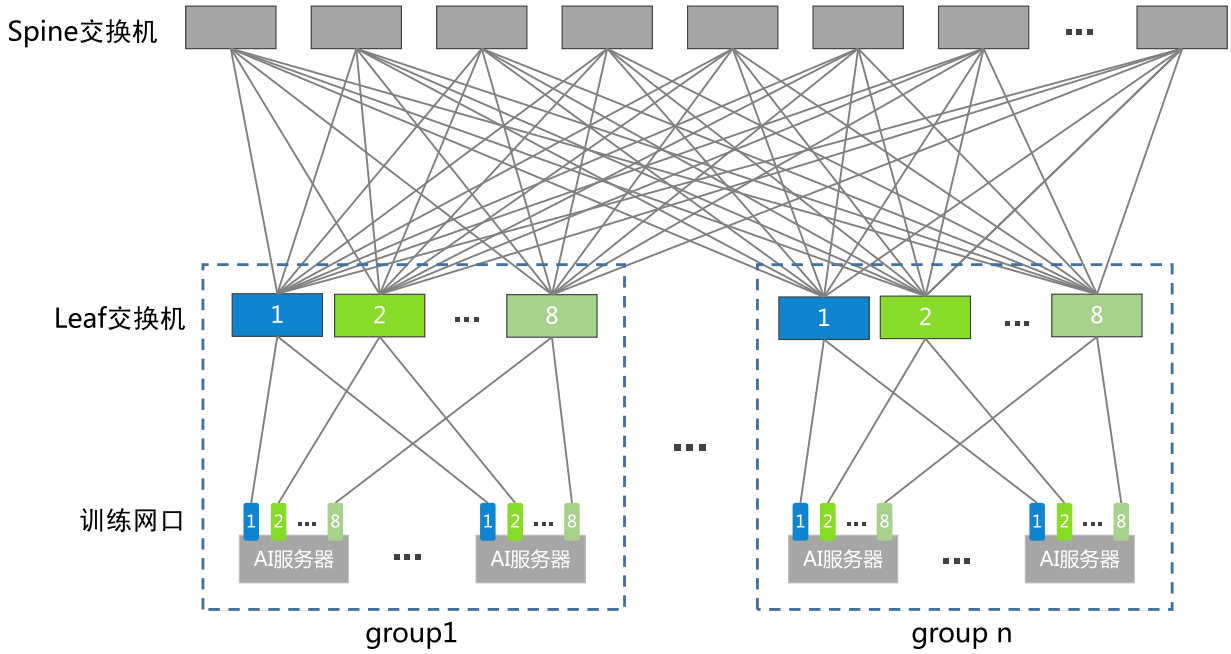

图 3 Spine-Leaf 两层组网

胖树（Fat-Tree）组网由 Leaf 交换机、Spine 交换机和 Core 交换机组成，如图 4 所示。每 8 台 Leaf 交换机和下挂的 AI 服务器做为一个 group，8 台 Leaf 交换机又和上面 N 台 Spine 交换机组成一个 pod，胖树组网以 pod 为单位进行扩展。在胖树组网中，Spine 交换机和 Leaf 交换机之间采用 Fullmesh 全连接，所有 Spine1 都 Full-Mesh 连接至第一组 Core，所有 Spine2 都 Full-Mesh 连接至第二组 Core，依次类推。Spine 交换机和 Leaf 交换机上下行收敛比都为 1:1。

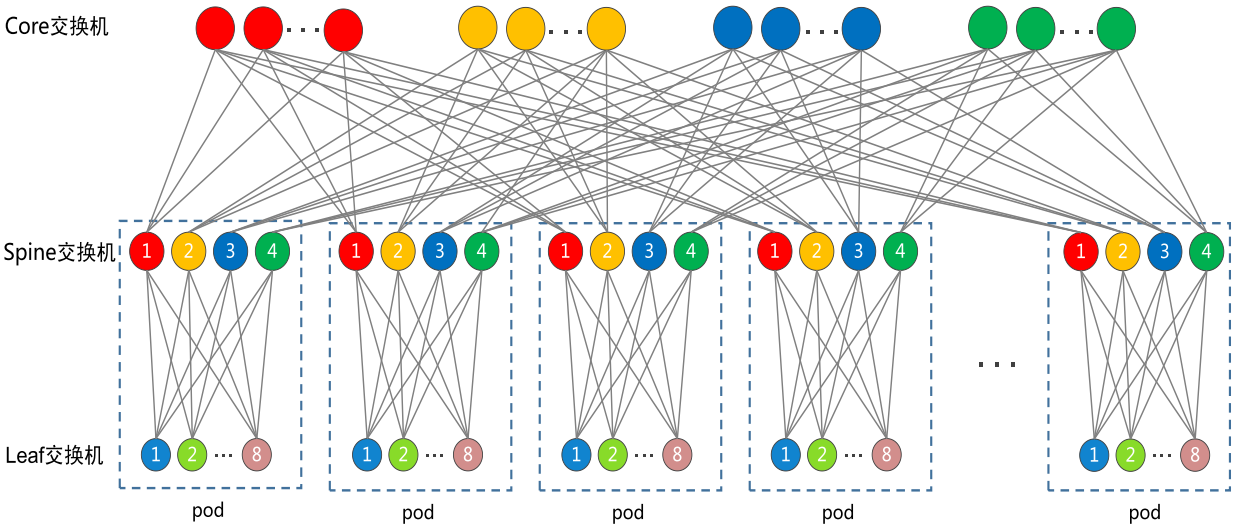

图 4 胖树组网

4.3.2 零丢包无损网络

分布式高性能应用的特点是“多打一”的 Incast 流量模型。对于以太交换机，Incast 流量易造成交换机内部队列缓存的瞬时突发拥塞甚至丢包，带来应用时延的增加和吞吐的下降，从而损害分布式应用的性能。AI 人工智能计算场景通常采用 RoCEv2 协议与 DCQCN（Data Center Quantized Congestion Notification）拥塞控制机制相互配合实现零丢包无损网络。

DCQCN 要求交换机对遇到拥塞的报文进行 ECN（Explicit Congestion Notification）标记，传统方式的 ECN 门限值是通过手工配置的，这种静态的 ECN 水线无法适配所有的业务流量模型；水线配置低了，频繁进行 ECN 通告，网络吞吐上不来；水线配置高了，可能导致频繁触发 PFC（Priority-based Flow Control），影响整网的其他业务流量。因此建议在参数面网络和数据面网络里部署动态 ECN 技术，通过 AI 算法，根据网络业务流量模型，计算出对应的 ECN 水线配置，达到在保证吞吐的同时，尽量维持较低的队列时延，让网络的吞吐和时延达到最佳平衡。

无论是静态 ECN 还是动态 ECN，本质上都是被动拥塞控制机制，通过反压源端降低报文发送速度来保证网络无损，实际上并没有达到提升吞吐率效果，反而降低了 GPU 利用率。因此，中国移动提出 GSE 技术，通过全局动态的主动授权机制，从根本上最大限度消除网络拥塞，减少网络设备队列资源的开销，降低模型训练任务的长尾时延，突破 RoCEv2 性能瓶颈。

4.3.3 高吞吐网络

AI 人工智能计算场景的流量特征是流数少、单流带宽大。传统的 ECMP（Equal Cost Multi Path）是基于 5 元组的逐流 HASH，在流数少的时候极易出现 HASH 不均的情况，建议使用端口级负载均衡技术或算网协同负载均衡技术代替传统的 ECMP。

端口级负载均衡部署在 Leaf 交换机上，以源端口或目的端口作为数据流均衡的影响因子，在一个端口组内将归属于不同端口的数据流均衡到本地出端口集合上，消除传统基于五元组哈希的不确定性。

除此之外，还可以在参数网络里部署算网协同负载均衡技术，AI 调度平台把任务信息通知给网络控制器，网络控制器结合已经建立的整网拓扑信息，进行整网最优转

发路径计算，计算完成后自动生成路径并动态下发到网络设备，实现多任务全网负载均衡。使网络吞吐可以达到 95%以上，接近满吞吐。

4.3.4 高可靠网络

超万卡集群中网络作为业务流量的调度中枢，其稳定性决定着整个集群的运行效率。在典型的 CLOS 组网中，交换机之间都有多条路径，当一条链路出现故障的时候，通过感知端口状态、路由收敛、转发路径切换等操作，完成流量从故障链路到备用链路的收敛。但是这个时间一般在秒级。然而在 AI 场景里面，每次通信时间在毫秒级别，秒级时间内正常情况下已完成了多轮通信。如果依靠传统的路由收敛方式，将极大的影响 AI 计算效率。

DPFR（Data Plane Fast Recovery）技术在此场景下，可以做到毫秒级收敛，提供基于数据面的本地快收敛或远程快收敛。特性包含故障快速感知，故障本地快速收敛，故障通告生成、接收和中继处理，故障远程快速收敛和表项老化处理。针对关键应用，尽量做到应用无感知的故障快速收敛效果，即在链路故障发生时业务性能无明显下降。

4.4 高容错高效能平台技术

智算平台是智算中心承载模型训练、推理和部署的综合性平台系统，在智算中心技术体系架构中承担着重要的角色，对算力基础设施进行统一纳管、调度、分配和全生命周期管理，主要包含对计算、存储、网络等 IaaS 资源的云化管控，在此基础上通过云原生的容器底座能力，实现智算资源纳管分配、AI 任务作业调度、拓扑感知调度、训练全链路监控等满足智算业务的核心需求。

随着模型参数量和数据量的激增，训练所需的单集群规模来到万级，但是智算平台的性能通常不能随着算力线性增长，而是会出现耗损，因此大模型训练还需要高效的算力调度来发挥算力平台的效能。而这不仅需要依赖算法、框架的优化，还需要借助高效的算力调度平台，根据算力集群的硬件特点和计算负载特性实现最优化的算力调度，来保障集群可靠性和计算效率。针对以上问题，业界多以断点续训、并行计算优化、智能运维等作为切入点，构建高容错高效能智算平台。

4.4.1 断点续训高容错能力

大模型训练面临的困难在于确保训练任务能够持续进行而不中断。在训练过程中，可能会遇到硬件故障、软件故障、网络故障以及其他故障。这种频繁中断导致的训练进度的损失对于耗时又耗资源的大模型训练来说是难以承受的，需要具备自动故障检测和训练重启。当前业界大模型训练主要容错方案依赖于训练过程中周期性保存checkpoint，故障发生后从上一个周期性 checkpoint 重新启动训练。

基于平台的运维监控能力，可以实现对超万卡集群的软硬件故障检测和预警，但是当故障发生且导致模型训练中断或停止时，通常需要人工介入排查故障信息，隔离故障并重新触发容器 pod 资源调度，重新完成并行训练中集合通信的初始化，重新加载基于中断前最近一次保存的 checkpoint 信息，最后经历算子库的重新编译，完成训练任务的继续。图 5 为典型的断点续训流程：

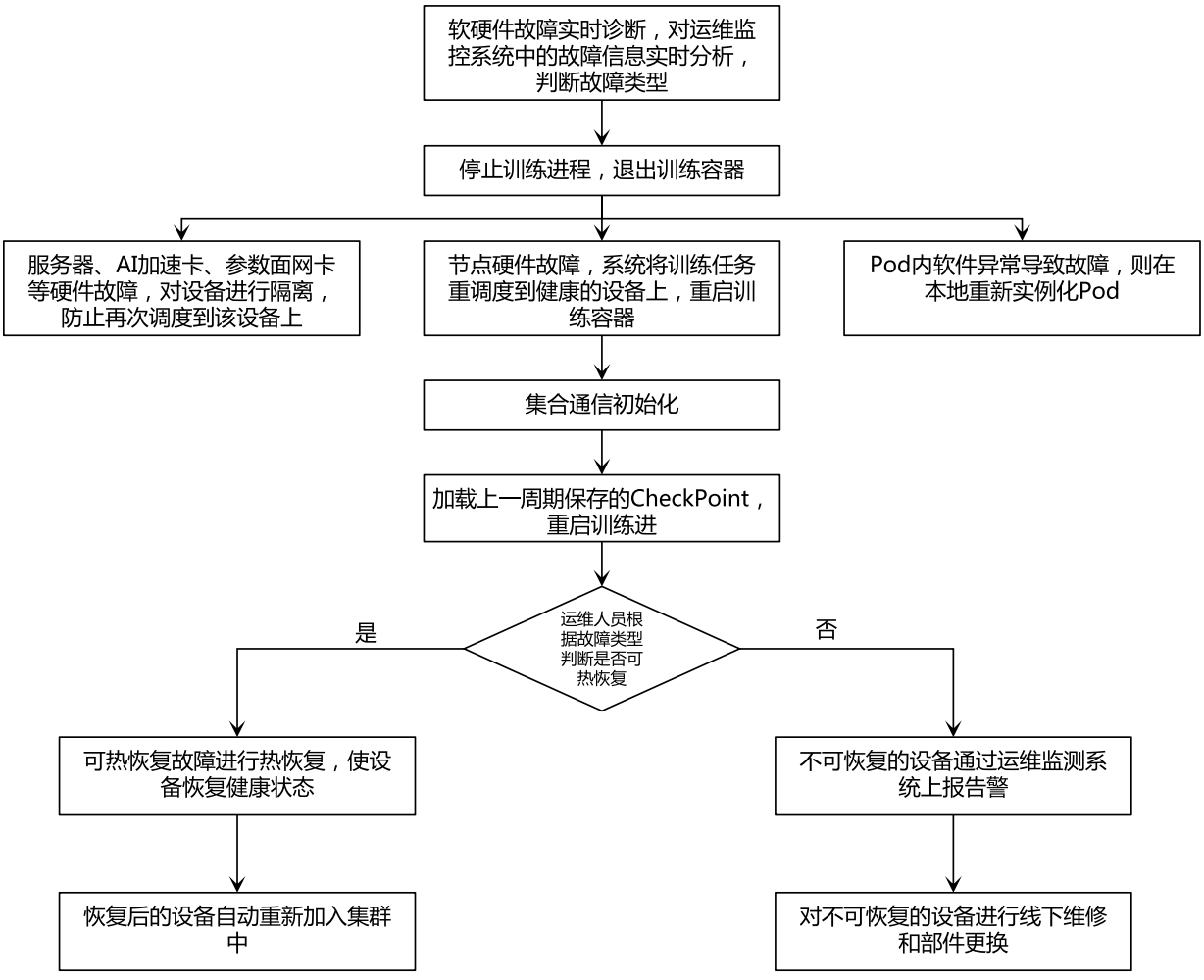

图 5 断点续训流程

在断点续训过程中，checkpoint 是模型中断训练后恢复的关键点，因此 checkpoint 密集程度、保存和恢复的性能尤为重要，checkpoint 本身的耗时与模型的大小成正比，当模型参数达到百亿甚至千亿时，checkpoint 的时间开销通常在几分钟到十几分钟之间。此时，训练任务需要暂停，使得用户难以频繁进行 checkpoint 操作，因此为保证训练效率，会适当拉长 checkpoint 保存周期。然而，一旦发生中断，之前损失的迭代次数在恢复时需要重新计算，需要花费更长的时间。

为解决该问题，需要尽量降低 checkpoint 流程的开销，既能大幅降低训练暂停时间，也能支持高频的 checkpoint 来减少容错时浪费的迭代步数。业界通常采用 checkpoint 多级存储的方式，构建基于更高 IO 性能的内存介质构建存储系统，相比于磁盘或者网络文件存储系统，checkpoint 在内存空间的保存可以大幅缩短训练暂停等待时间。同时，结合业务需求定期地将 checkpoint 异步写入到持久化的存储系统中，异步流程不干扰正常的训练。当发生故障导致训练任务重启时，由于内存系统中的 checkpoint 数据并未丢失，新启动的训练进程可以直接读取内存系统中的 checkpoint 数据来加载模型和优化器状态，从而省去了读取网络存储系统的 IO 开销。

断点续训基于多级 checkpoint 存储、软硬件协同优化以及全栈系统级容错，实现训练任务分钟级恢复，在技术价值方面，实现故障检测、故障隔离、资源重调度、训练任务恢复无人工全流程自动化；在商业价值方面，作为智算平台关键特性提供给模型开发者使用，保障大模型训练任务长期稳定运行，提升用户满意度。

4.4.2 分布式并行计算优化

超万卡集群中分布式并行训练框架[8]、[9]是标准配置，即在大规模算力资源池上搭建用于并行处理深度学习模型分布式训练任务的工具集合，其将训练任务划分为多个子任务，通过在多台计算机上并行执行，实现高效、可靠和快速的分布式大模型训练，提高模型的准确性和效率。

超万卡集群因节点数众多、资源类型不统一、数据量大、网络环境复杂，给大模型训练带来了许多挑战。

挑战一：实现大规模的高效率训练。Model FLOPs utilization (MFU)是实际吞吐量与标称最大吞吐量之比，是评估模型训练效率的通用指标，可以直接反映端到端的训

练效率。为了训练大模型，需要将模型分布为多个 GPU 上，并且 GPU 之间需进行大量通信。除了通信之外，如操作符优化、数据预处理和 GPU 内存消耗等因素对 MFU 也有着显著影响。

挑战二：实现训练的高稳定性，即在整个过程中保持高效率训练。在大模型训练中，稳定性十分重要，失败和延迟虽是大模型训练中的常态，但其故障成本极高，如何缩短故障恢复时间至关重要。

目前分布式并行框架在进行模型训练时流程如下：

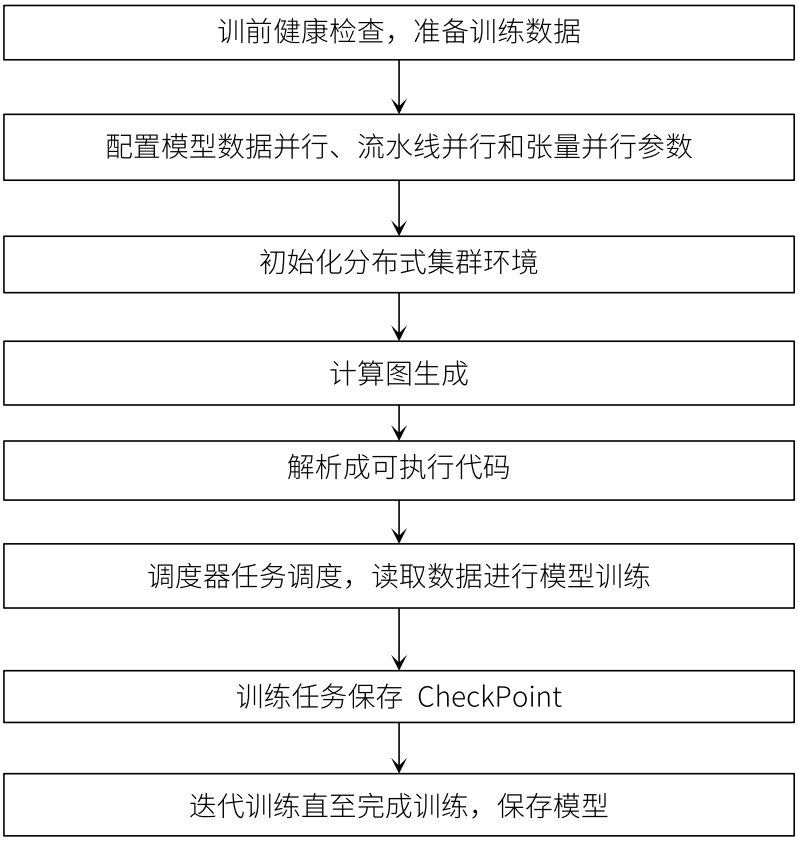

图 6 分布式并行训练流程

其中，每个步骤都涉及到影响模型运行效率的问题。针对如上步骤，超万卡集群分布式框架需针对以上流程进行优化，且支持更多类型的模型加速训练技术，如自动并行方案生成、自动触发计算图优化，数据流水线管理等。

- 自动 3D 并行策略：支持基础数据并行、模型并行及流水线并行的一种或多种组合形态。
- 自动并行方案生成：根据模型结构及参数量、现有硬件资源拓扑情况、网络带

宽等信息，以通信代价最小为目标，自动生成配置模型训练过程的最优 3D 并行参数组合。

- 自动计算图优化：计算图作为连接深度学习框架和前端语言的主要中间表达，被目前主流框架如 TensorFlow 和 PyTorch 所使用或者作为标准文件格式来导出模型。计算图的执行效率极大程度上影响代码执行效率。构建高效算子库，预置算子融合库，设计子图替换规则，以计算图在节点间的计算通信信息为输入，尽量使每个节点的负载做到均衡，在编译阶段，触发自动计算图优化。
- 数据流水线优化：数据预处理和训练数据加载阶段 GPU 空闲，可采用数据并行处理、数据分布式加速缓存等技术，优化提升 GPU 利用率，提升模型训练效率。
- 显存优化：显存优化方法通常包括代码优化、梯度累计、存储格式优化、半精度训练等，通过综合使用多种显存优化方法，降低显存消耗，提高大模型训练的稳定性和可扩展性。

4.4.3 超万卡集群智能管控

随着智算集群规模不断扩大，集群运维管控与集群应用之间的矛盾日益凸显。随着单集群的 AI 加速卡数量从从千级增长到万级，相应的故障范围扩大超过 10 倍。典型故障范围从单服务器单卡迅速扩散到算网存多域全栈。为彻底改变这一运维困境，亟需引入新的运维理念和技术，以集群全链路可视化监控、故障快速定位和运维侧快速修复为原则来建设新的集群计算智能运维系统。

超万卡集群智能运维系统需要具备算、网、存协同管理的端到端系统运维管理能力，包括计算设备、网络设备、存储设备、光模块设备管理、控制以及分析等全生命周期运维管理能力，提升训练效率、降低训练成本，实现大模型训的快、训的稳、训的好。

新的集群智能运维管理系统从底层建设开始应具备统一的容器化平台与公共技术底座。系统南向的实现应采用统一的采集框架，统一对被管理的计算、存储、网络、光模块单元进行资源、性能、告警、日志、拓扑等信息的采集，并存放到集中数据底座中。运维系统应构建公共的服务及数据底座，为整个系统提供基础服务及数据存储能力，并基于容器化平台与公共技术底座构建基础的公共服务，提供资源管理、服务

编排、监控、作业运维等功能，实现对万卡智算集群的智能运维服务。

集群计算智能运维管理系统在实际业务布局中应具备与 AI 作业任务密切相关的能力。一般情况下应具备 AI 作业路径可视功能、环境健康检查功能、AI 训练作业故障诊断、集群环境管理、集群资源管理、服务器管理以及监控分析等能力。

- **作业路径可视**

作业路径需支持展示与训练作业相关的资源视图，包括参数面交换机、智算服务器、AI 加速卡和链路，提供路径拓扑可视化运维管理。用户可通过作业路径功能查询管理与训练作业相关的资源和关键数据。作业路径可视的业务范围包括训练作业关联资源管理，支持填写训练作业 ID，查询关联资源，管理作业分布资源并查看相关 Issue 和 KPI 指标。支持可视化展示作业链路的设备运行状态，动态评估任务路径的健康状况。

- **环境健康检查**

环境健康检查包括集群环境健康检查和作业运行前环境检查两种类型，均与 AI 作业训练的环境准备相关。集群环境健康检查：对集群环境进行全面检查，规避软硬件问题。对集群软硬件、环境配置及性能进行健康评估，并输出完整报告。作业运行前环境检查：对作业运行所需资源的健康、一致性、性能等进行检查，提高作业运行成功率。

- **故障诊断**

对执行失败的作业进行智能故障诊断定界，分析全链路影响因子，基于作业运行环境日志进行诊断，覆盖常见软件栈报错，同时对计算、网络、存储域的告警、故障进行时空关联分析，实现多种典型故障的实时诊断。基于 AI 训练场景搭建故障知识库，提供快速精准的故障知识检索能力。

- **集群资源管理**

提供作业集群管理能力，包括集群名称、状态、计算节点个数等信息。 支持对集群信息的配置和修改能力，提供运维管控面的集群相关元数据管理。

- **设备管理**

提供集群内计算、网络、存储、光模块等全量设备的静态数据录入、采集和管理

能力，能够快速检索查看设备的详细信息，提升 AI 集群资源的管理效率。支持将服务器设备添加到智能运维管理系统进行统一管理，监控资源、告警、性能数据，提升管理效率。

- **监控分析**

基于静态数据、性能数据以及日志数据，提供软硬件、作业等多维度的性能分析，动态评估和监控软硬件、作业的运行状态，提供监控分享告警、日志分析与检索，快速识别关键信息，提升日常运维效率。

- **监控大盘**

实时呈现集群资源使用和运行状况、训练任务得执行情况等相关监控数据，动态展现一段时间内性能的变化情况，实现对关键监控数据集中显示的功能，及时、全面、快捷的地掌握集群的运行状况。

4.5 新型智算中心机房设计

面向高密度高能耗智能算力发展，对于部署超万卡集群的新型智算中心来说，需要在确保智能计算设备安全、稳定、可靠地运行的前提下，具备高效制冷、弹性扩展、敏捷部署、绿色低碳等特征，并实现智能化运维管理。新型智算中心机房的关键要素如下：

4.5.1 高效制冷

智算中心催生了海量算力需求，芯片 TDP 不断攀升，风冷难以散热，同时也带来总功耗不断增加，散热和能耗成为智算中心迫在眉睫的问题，液冷具有散热效率高以及支持更高功率处理器的优势，近年得到了快速发展，可推动扩大解耦型冷板液冷或单相浸没液冷技术应用范围及推动交换机等网络设备应用液冷，解决高密服务器散热的同时降低智算中心整体能耗，另外解耦冷板液冷可以实现基础设施侧与 IT 设备侧解耦，实现智算业务快速弹性部署。

4.5.2 弹性供电

智算中心具有高密度、负载率波动大的特点，需弹性供电以适配不同计算任务需求。供电系统将采用大容量、模块化高效不间断电源，形成电力资源池，以每列智算

机架为颗粒度，预留高密机架和普通密度机架互弹条件，提高系统效率和灵活性；采用末端小母线供电（或列头柜预留出线回路）的机柜供电方案，提升末端供电的灵活性。对于未来超高功率的智算机柜，采用放射式供电、高集成度电力模块等方案，节省占地，提升平面布局的灵活性。

4.5.3 敏捷部署

智算业务需求短时爆发，敏捷部署的智算中心成为刚需。新型智算中心规划建设时，可采用一体化电源系统、预制集成泵站模式、集装箱式智算中心、模块化智算中心等预制模块化建造技术，缩短工程交付周期，实现快速部署。

4.5.4 绿色能源应用

新型智算中心应积极应用绿色能源技术，实现低碳零碳算力和可持续发展。新型智算中心应结合园区选址特点与周边环境条件，因地制宜部署分布式光伏、风力发电等系统，实现清洁能源的就地生产与消纳；通过电力交易、绿色证书交易等模式采购可再生能源电力，提升绿色能源使用比例。随着氢能应用技术的发展，智算中心可内逐步规模化应用氢燃料电池。

4.5.5 智能化运维管理

借助大数据、AI 技术、数字孪生等技术，构建新型智算中心的智能运维管理体系。运用 AI 算法预测设备故障、优化能源使用、智能调度资源，实现主动运维、精准运维。通过机器学习、大数据分析等技术，对智算中心的运行数据进行深度挖掘，提升故障诊断、性能调优、容量规划等方面的决策准确性与效率。

第五章：未来展望

随着数据规模的持续扩大、集群能力的不断增强以及大模型应用的日益丰富，对新型智算底座的升级提出了更高的要求。面对未来，我们呼吁在超节点、跨集群训练、软件框架等领域实现技术突破，以强化智算基础设施能力。与此同时持续探索存算一体、光子芯片等先进技术领域与智算中心的结合，为下一次信息变革奠定基础。

1）**引入超节点，拓展 Scale up 能力：**随着大模型的进一步发展，单纯通过 Scale out 扩展更多张 AI 卡已经无法满足万亿、数十万亿大模型的训练需要，算力形态将通过 Scale up 发展到超节点架构，突破传统单机 8 卡，通过内部高速总线将 AI 芯片互联，一台超节点即可实现万亿参数训练和实时推理，未来超节点将成为智算基础设施的重要组成部分。面向未来数万乃至数十万卡超大规模组网、高速总线无收敛互连、统一内存语义互访、数十乃至数百 MW 级供电散热等等，仍需重点攻克。

为了支持 scale up 卡间互联能力，中国移动提出一种创新的互联架构——全向智感互联系统（Omnidirectional IntelligentSensing Express Interconnect Architecture，简称 OISA，音译“欧萨”），旨在为 GPU 间南向通信提供优化的连接方案。OISA 将基于对等通信架构、极简报文格式、高效物理传输和灵活扩展能力等设计理念，构建一套可以支持百卡级别的 GPU 高速互联系统，在支持卡间交换拓扑的同时，通过对电接口、聚合技术、报文格式进行优化，提高 GPU 之间的数据传输效率。OISA 将在物理层、链路层、事务层等方面进行系统性重构，为大规模并行计算和 AI 应用构建一个高效、可靠的互联能力，以支持非平面布局的多维互联，打破传统服务器内连接限制，实现高效数据协同。

2）**大规模逻辑集群，突破传输距离限制，探索跨节点互联网络技术：**随着模型参数量、算力资源需求十倍速增长，驱动智算中心组网规模向万卡级，甚至是十万卡级演进。智算中心因机房空间、供电等基础设施限制，不可避免出现同园区跨楼宇部署及小局点短距互联实现逻辑大集群的需求。网络传输距离拉远会增加传输时延以及对传输设备的无损缓冲提出了更高的要求，相应也会影响集群有效算力，需要从工程上和科学上进一步研究和验证影响性和优化方案。

3）**软件框架技术方面，提升自动化能力和训练效率：**超万卡集群下模型规模和数

据集复杂度提升，需要在软硬件、算法、网络等方面持续创新，聚焦于自动化、跨平台支持、大规模模型训练、跨集群训练、边缘训推等方面不断优化完善，实现高效、可靠和快速的深度学习模型训练，提高模型的准确性和训练效率，降低用户开发大模型的使用门槛和资源开销，提供更加高效、易用的模型训练工具。

4）**潜在换道超车技术方面，突破摩尔极限，探索下一代芯片设计和应用范式：**大模型的发展给传统冯诺伊曼计算体系结构带来了功耗墙、内存墙和通讯墙等多重挑战。未来需探索从存算一体、光子芯片等领域突破现有 AI 芯片设计和应用范式，一方面大力推动存算一体在大模型推理场景应用，推进先进制程支持，加速存算一体技术在大模型芯片和大规模应用；另一方面是利用好光子芯片在传输速度、低功耗等方面的技术优势，探索未来与智算产业和 AI 生态的结合方式。

缩略语列表

缩略语	英文全称	中文解释
AI	Artificial Intelligence	人工智能
C2C	Chip-to-Chip	芯片到芯片
CDU	Coolant Distribution Unit	冷量分配单元
CPO	Co-Packaged Optics	光电共封装
DCQCN	Data Center Quantized Congestion Notification	数据中心量化拥塞通知
DP	Data Parallel	数据并行
DPFR	Data Plane Fast Recovery	数据面故障快速恢复
DPU	Data Processing Unit	数据处理单元
DSA	Domain Specific Architecture	特定领域架构
ECMP	Equal Cost Multi Path	等价多路径
ECN	Explicit Congestion Notification	显示拥塞通知
FC	Fully connected	全互联拓扑
GSE	Global Scheduled Ethernet	全调度以太网
HBM	High Bandwidth Memory	高带宽内存
IaaS	Infrastructure as a Service	基础设施即服务
IB	InfiniBand	“无限宽带”技术
IOPS	Input/Output Per Second	每秒输入/输出操作次数
MFU	Model FLOPs Utilization	集群有效算力
MoE	Mixture of Experts	专家并行
MP	Model Parallel	模型并行
MTBF	Mean Time Between Failure	平均故障间隔时间
MTTR	Mean Time To Repair	平均修复时间
NICC	New Intelligent Computing Center	新型智算中心
NFS	Network File System	网络文件系统
NPO	Near Packaged Optics	近封装光学
OISA	Omnidirectional Intelligent Sensing Express Interconnect Architecture	全向智感互联
P2P	Point to Point	点对点
PFC	Priority-based Flow Control	基于优先级的流量控制
POSIX	Portable Operating System Interface	可移植操作系统接口
PP	Pipeline Parallel	流水线并行
PUE	Power Usage Effectiveness	电能利用效率
RDMA	Remote Direct Memory Access	远程直接数据存取
RoCE	RDMA over Converged Ethernet	融合以太网承载 RDMA
S3	Sample Storage Service	简单存储服务
TTA	Time to Accuracy	模型训练至预定精度的时长
UEC	Ultra Ethernet Consortium	超以太网联盟

参考文献

[1] Kaplan J , Mccandlish S , Henighan T ,et al.Scaling Laws for Neural Language Models[J]. 2020.DOI:10.48550/arXiv.2001.08361.

[2] Shazeer N , Mirhoseini A , Maziarz K ,et al.Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer[J]. 2017.DOI:10.48550/arXiv.1701.06538.

[3] Touvron H, Martin L, Stone K, et al. Llama 2: Open foundation and fine-tuned chat models[J]. arXiv preprint arXiv:2307.09288, 2023.

[4] 中国移动 NICC 新型智算中心技术体系白皮书，中国移动,2023

[5] Jiang Z, Lin H, Zhong Y, et al. MegaScale: Scaling Large Language Model Training to More Than 10,000 GPUs[J]. arXiv preprint arXiv:2402.15627, 2024.

[6] B. Kim et al., "The Breakthrough Memory Solutions for Improved Performance on LLM Inference," in IEEE Micro, doi: 10.1109/MM.2024.3375352.

[7] 全调度以太网技术架构白皮书，中国移动研究院

[8] Shoeybi M, Patwary M, Puri R, et al. Megatron-lm: Training multi-billion parameter language models using model parallelism[J]. arXiv preprint arXiv:1909.08053, 2019.

[9] Rasley J, Rajbhandari S, Ruwase O, et al. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters[C]//Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2020: 3505-3506.