CAICT 中国信通院

集智 专题报告

高质量大模型基础设施研究报告

（2024 年）

中国信息通信研究院人工智能研究所

2025年1月

版权声明

前　言

随着大模型技术的飞速发展，模型参数量急剧增长，模型能力持续增强，智能应用百花齐放。基础设施的可用性决定了大模型研发及服务的效率，大模型服务的可用性又决定了智能应用的服务质量。在此背景下，高质量大模型基础设施成为推动大模型应用落地的关键要素。

目前，大模型基础设施普遍面临可用性低、稳定性差等问题，亟需从计算、网络、存储、软件和运维等多层面协同优化。在同等计算资源条件下，通过多系统协同优化提升基础设施可用性，不仅可以提高大模型开发效率和服务能力，还能有效降低应用成本，加速大模型规模化落地。

本报告聚焦大模型基础设施的五大核心能力领域：计算、存储、网络、开发工具链和运维管理，系统梳理了大模型发展对基础设施提出的新需求，剖析基础设施发展的关键技术，并提出体系化评价指标。同时，通过分析业界典型实践案例，为企业建设高质量大模型基础设施提供参考。

展望未来，大模型基础设施将与大模型一起迭代升级，并为大模型的规模化应用提供有力支撑。本报告力求为相关领域的研究与实践提供参考，但难免有不足之处，恳请各位专家和读者不吝指正。

目 录

图 目 录

表 目 录

一、大模型基础设施概述

大模型技术作为人工智能领域的突破性进展，正迅速推动各行各业的智能化转型。随着参数量的增长，大模型展现出强大的理解能力和复杂数据处理能力，在金融、医疗、政务等行业的应用日益广泛。然而，参数量的增加也给大模型落地带来了巨大的挑战，提高大模型基础设施能力，满足大模型全生命周期对基础设施的新需求成为首要任务。

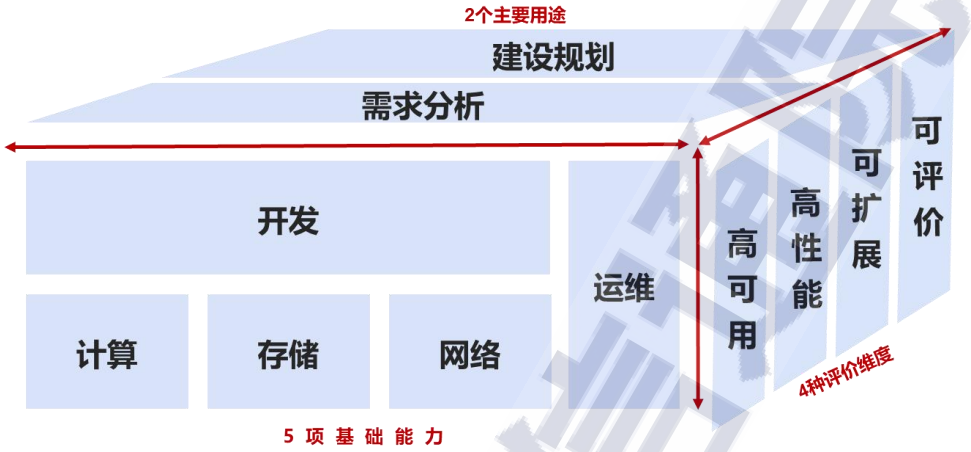
（一）大模型基础设施概念与特性

大模型基础设施是指支持大规模人工智能模型（如大语言模型、多模态大模型等）训练、部署和应用的硬件和软件资源的集合，它包括高性能计算、海量数据存储、高速网络连接资源，以及相应的软件框架和工具链，旨在为大模型的开发、训练和推理提供高效、可靠和可扩展的支撑环境。大模型基础设施整体架构如图 1 所示。

来源：中国信息通信研究院

图 1 大模型基础设施架构图

大模型全生命周期要求大模型基础设施具备高可用、高性能、可扩展、可评价等特性。大模型基础设施能力如图 2 所示。



来源：中国信息通信研究院

图 2 大模型基础设施能力矩阵

（1）高可用：稳定的大模型业务支撑能力

高可用是指在提高大模型基础设施平均无故障运行时间（Mean Time Between Failures，MTBF）的同时考虑更短的平均故障定位时间（Mean Time to Identify，MTTD）和平均故障恢复时间（Mean Time To Recovery，MTTR），综合考虑存储、运维、开发软件等维度。

可用度是指大模型基础设施集群在一定时间内提供正常服务的时间占总时间的比例，通常用百分比表示。数据显示，当前集群可用度普遍低于 50%。Meta 50000+卡训练任务[1]和 OpenAI GPT-4 25000 卡训练任务集群算力可用度在 30%～40%之间，英伟达 Megatron-LM 和微软 MT-NLG 10000+卡训练任务的集群算力可用度在 40%～50%

[1]Meta. "Building Meta's GenAI Infrastructure". https://engineering.fb.com/2024/03/12/data-center-engineering/building-metas-genai-infrastructure/.

之间，字节跳动万卡集群 MegaScale 集群算力可用度仅达 55.2%[2]。大模型基础设施的可用度仍有较大提升空间。

平均无故障运行时间是全系统维度的考量，指大模型基础设施运行时相邻两次故障之间的平均工作时间，也称为平均故障间隔。平均故障定位时间是运维维度上的考量，指大模型作业运行时，基础设施集群出现故障到故障首次被发现的平均时间，关注的是故障定位效率。平均故障恢复时间是存储、开发软件、运维等维度的考量，指大模型基础设施发生故障后修复所需的平均时间，关注故障恢复效率。

（2）高性能：高效的大模型业务运行能力

高性能是指提高大模型基础设施的算力供给能力。算力供给能力即“大模型基础设施算力规模”乘以“算力利用率”，综合考虑计算与开发软件等维度。

算力规模和硬件算力利用率是计算维度的考量，算力规模指大模型基础设施理论计算规模，计算方式为“单节点算力规模”乘以“节点数”，理论算力规模数值越大，代表潜在的计算能力越大。硬件算力利用率（Hardware FLOPs Utilization，HFU）是指考虑重计算后，模型一次前反向计算消耗的矩阵算力与机器算力的比值。硬件算力利用率越高，代表资源利用越充分。

（3）可扩展：资源需求与技术发展的共同选择

可扩展指的是大模型基础设施在负载增加时，通过增加资源维持或提高性能的能力，在具备扩建能力的同时，兼具技术兼容的特性。

[2]Jiang, Ziheng, et al. "MegaScale:Scaling large language model training to more than 10,000 GPUs." 21st USENIX Symposium on Networked Systems Design and Implementation (NSDI 24). 2024.

在资源需求上，可扩展性体现在既有基础设施的扩建和有效利用上。随着企业大模型业务需求的不断增长，在成本控制的背景下，企业希望充分利用既有基础设施，对现有基础设施升级改造同时预留二次升级改造的接口，以适应不断发展的大模型业务需求。

在技术发展上，可扩展性体现在对计算、网络、存储、开发、运维等技术的“前”向接口兼容和“后”向更新迭代上。计算软件、网络协议、存储软件、开发平台和运维平台等应支持不同品类、不同协议、不同技术路线的硬件产品，以实现大模型基础设施的可扩展。

（4）可评价：多角度反映大模型基础设施应用成效的多元评价

可评价是指面向大模型应用场景，大模型基础设施可通过完整、有效的评价体系反映其赋能成效。

当前大模型基础设施评价体系存在评价维度单一等问题，亟需扩充评价维度，以便更系统、更全面、更深入地反映大模型基础设施赋能效果。一是明确大模型基础设施的评价指标，确定需要评估的特性或参数，如计算能力、存储性能、网络带宽、可靠性、可扩展性等。二是建立评价方法，采用适当的工具和技术，收集和分析相关数据，以评估大模型基础设施的整体能力。三是确定用于评价的数据采集与分析方法，通过检测、测试或模型模拟的方式，获取大模型基础设施运行过程中的数据，分析得出评价结果。

（二）大模型基础设施现状

技术方面，AI 存储能力提升，进一步提高基础设施可用度。橡树岭实验室应用戴尔、DNN 公司等的新一代 AI 存储，显著提升了数

据存取速度，华为、清华大学 MADSys 实验室联合开发的高密高性能 AI 存储获得 MLPerf Storage 基准评测第一名，为大模型基础设施的存储优化提供技术支撑。**网络技术不断涌现，大模型基础设施通信效率提升。**中国电信提出了新的 RDMA 端到端拥塞控制机制[3]，该技术不仅可以有效提升智算中心网络通信效率，提高整体系统训练效率，同时还能够降低训练成本，有效提升国产化网络技术竞争力。

产业方面，科技大厂已形成完整的大模型基础设施生态。亚马逊、微软、谷歌等厂商在大模型基础设施领域占据领先地位，已形成较为完整的生态系统。如亚马逊、微软等已实现算力、平台、模型、软件的垂直整合，统一对外提供服务。百度、阿里、腾讯、华为等科技巨头纷纷加大在大模型基础设施方面的投入，均已形成涵盖“AI 计算平台+AI 开发平台+大模型”的全产业生态。如百度智能云提出“打造大模型的新质基础设施”、商汤提出“AI 基础设施新范式-商汤大装置”等。

政策方面，各国家加大资金投入力度促进大模型基础设施发展。2024 年 9 月，拜登政府宣布将启动“AI 数据中心基础设施工作组”，旨在协调政府各部门，加速 AI 数据中心建设，以满足 AI 日益增长的电力和算力需求。同时美国能源部将策划一套资源包（包括贷款、赠款、税收抵免和技术援助），帮助 AI 数据中心所有者和运营商获得清洁、可靠的能源解决方案。2024 年 6 月德国发布《人工智能计算基础设施行动计划》，目的是为工业界的人工智能开发人员提供具有

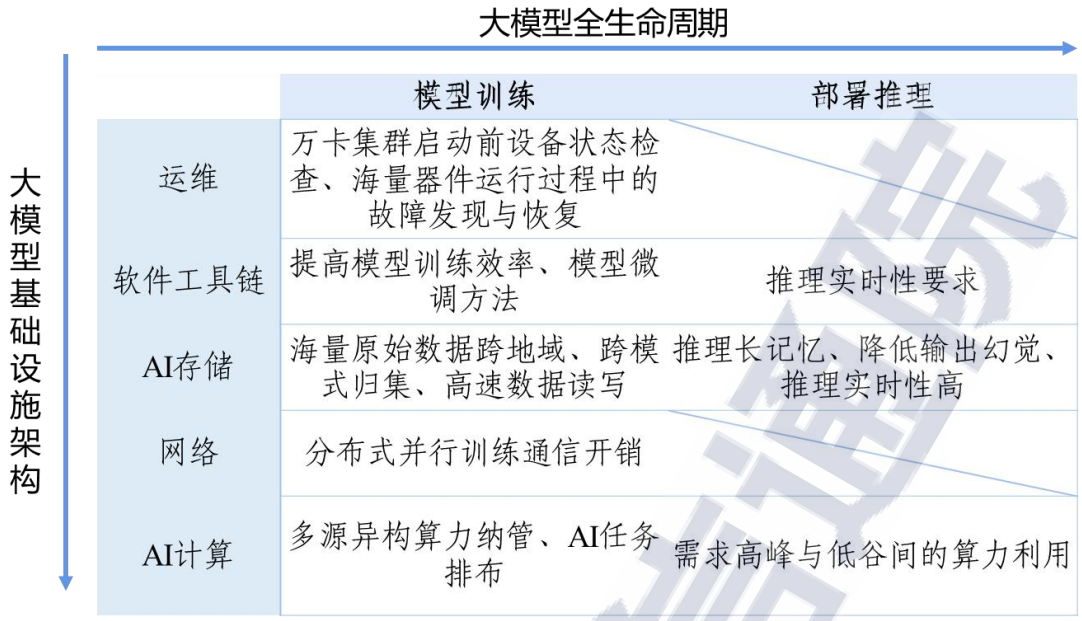
[3] 车碧瑶等."运营商大模型硬件基础设施创新及 RDMA 流量控制技术研究." 信息通信技术与政策 002(2024):050.

国际竞争力的计算能力。2024 年 9 月韩国发布《国家 AI 战略政策方向》，提出要扩大建设国家 AI 计算基础设施，计划以公私合资方式建设“国家人工智能计算中心”。我国相继出台《算力基础设施高质量发展行动计划》《关于深入实施“东数西算”工程加快构建全国一体化算力网的实施意见》等文件，将算力基础设施及公共数据资源纳入国家创新要素供给，指导智能基础设施有序布局，避免重复建设。

二、大模型基础设施挑战

当前，大模型基础设施集群可用度低，可利用算力难以随集群部署规模线性增长。英伟达、斯坦福大学和微软研究中心联合发表的论文[4]显示，算力规模增加的同时，集群可用度明显下降。大模型全生命周期对大模型基础设施提出新需求，如图 3 所示。大模型基础设施亟须通过算、存、网、软件、运维协同优化提升可用度。技术上，要求大模型基础设施具备高密度算存硬件、高性能无阻塞网络以及高并行度的通信和计算范式。

[4]D.Narayanan,et al, "Efficient Large-Scale Language Model Training on GPU Clusters Using Megatron-LM"SC21: International Conference for High Performance Computing,Networking, Storage and Analysis,St. Louis,MO,USA,2021,pp.1-14.



来源：中国信息通信研究院

图 3 大模型全生命周期对大模型基础设施的关键需求

（一）计算资源分配粗放，利用率低成为新难题

大模型参数由千亿向万亿发展，算力需求骤增。以 GPT-4 为例，其具有 1.8 万亿参数，训练数据约 13 万亿个 Token，训练算力约 2.15×10^{25}FLOPS，相当于在 2.5 万张 A100 加速卡上运行 90～100 天[5]。大模型对计算资源的需求日益扩增，要求计算资源分配高效合理。

训练任务排布不合理，资源碎片化严重。一是多用户环境中，不同用户的资源需求和使用模式互不相同，部分用户侧重大规模训练任务，部分用户仅运行小规模微调、推理任务。不同用户的任务只通过可用资源总量进行限制，极易导致资源分配混乱，导致资源碎片化。**二是**大模型整节点训练任务比例逐渐提升，但仍存在分布式训练任务、开发微调、推理服务等多任务场景混合的调度场景，优先保证大模型

[5] Dylan Patel and Gerald Wong. "Demystifying GPT-4: The engineering tradeoffs that led OpenAI to their architecture".https://www.semianalysis.com/p/gpt-4-architecture-infrastructure.

整节点训练任务，意味着对调度优化策略有着更高的要求。

推理请求波峰波谷现象明显，推理算力空闲产生资源浪费。一方面，业务在进行模型部署时会绑定固定的算力资源，可能出现多个 AI 推理任务抢占一张推理卡的情况，而其他推理卡还有空余资源，出现算力资源的浪费。**另一方面，**推理场景情况复杂，不同场景任务所消耗资源波峰、波谷差距较大，时间分布往往无明显规律，而推理服务往往有较高及时响应要求，意味着对推理系统优化和服务调度有更高要求。

围绕 AI 计算芯片设计的算力调度系统存在资源超额申请问题。基础设施的资源利用率和任务执行效率是由多维度资源共同决定。当前算力调度多将 AI 计算芯片作为影响任务性能表现的主要因素，忽略了基础设施中的 CPU、内存、网络等其他维度资源的影响。为保证计算任务顺利执行，用户在提交任务时通常会进行超额的资源申请。在调度系统中，超额申请的资源将被标记为已消耗，易导致后续任务因可用资源不足无法计算，产生计算资源浪费。

（二）海量数据处理低效，数据存储成为新瓶颈

数据总量和质量决定了大模型能力的上限。根据“尺度定律（Scaling Law）”，增加训练数据量，大模型训练效果会越来越好，GPT 系列的训练数据由 GPT-1 的 4.6GB 增长至 GPT-4 的约 40TB，Sora、Gemini 等多模态大模型发展带动训练数据需求十倍、百倍级增长。海量数据的准备效率和数据在全流程间的流转效率是影响大模型端到端生产成本的核心要素，AI 存储是解决数据归集时间长、数据处

理效率低、记忆时间短等问题的核心环节。

海量原始数据归集等待时间长。一是大模型训练所需的多模态数据需事先汇集。数据汇集需通过多地域、多渠道汇总，涉及数据中心、边缘设备等不同层级设备和不同协议之间的数据交互。据统计，大模型训练所需的 PB 级原始数据归集通常需耗时 3~5 周，占据整个大模型研发时长的 30%。**二是**采集的训练数据小文件居多，元数据管理困难。郑纬民院士论文显示[6]，任一模态数据集可能包含数亿甚至数百亿小文件，存储 100 亿小文件需管理 7TB 元数据。海量小文件的元数据处理时间长，要求文件系统既要扩展性好又要读取速度快。

训练数据预处理时间长。一是采集的原始数据无法直接用于模型训练。从多渠道收集的海量数据往往良莠不齐，直接用于模型训练会影响模型训练效果。为了获得高质量的数据样本，在模型训练前需要进行数据的预处理，如随机采样、数据解码、变换等操作。谷歌数据中心统计，大模型训练时间的 30%是用于数据预处理。微软分析了 9 种常见模型，数据预处理最多占用 65%的模型训练时间。**二是**数据读取网络开销大，现有方法通常以计算为中心，训练过程中待处理数据可能分散在多个节点，将需要处理的数据转移到计算节点时，读取远端节点的数据会产生极大的网络开销。

训练阶段检查点（Checkpoint）文件读写效率低。为提高大模型训练效率，实现断点续训，在训练过程中需要存储检查点文件。以国家超级计算无锡中心的神威平台 10 万卡规模训练万亿参数量模型为

[6]郑纬民. 分布式技术在大模型训练和推理中的应用[J]. 大数据, 2024, 10(5): 1-10.

例[6]，需保存近 12TB 的模型参数到检查点文件中，在未经优化的情况下，单次检查点文件写入需花费 3 小时，等待写入过程中的计算资源闲置，导致基础设施可用度不足。

大模型推理记忆时间短、存在输出幻觉。一是记忆时间短导致大模型进行长对话时无法精准理解用户需求，个性化体验差。存储推理过程中的长上下文及中间推理 token 成为提升大模型逻辑能力的重要优化方向。**二是**输出幻觉易导致大模型信任危机，阻碍大模型落地应用。引入外部知识库可将推理过程中的内容生成转换为基于既有数据的搜索或摘要问题，避免因模型内部数据不足或偏差产生的错误结论。外部知识库的引入要求 AI 存储具备高效的高维数据处理能力和复杂的查询操作能力。

（三）并行计算规模攀升，网络通信成为新阻碍

随着大模型参数的骤增，建设应用单体超大规模集群成为大模型基础设施的设计目标。领先的科技公司正积极部署万卡甚至十万卡规模的计算集群，大集群带来的网络互联挑战正成为大模型训练提速的新阻碍。

大模型训练通信开销大，多机多卡互联性能成为大模型训练瓶颈。以在超万卡集群训练 1.8 万亿参数的 GPT-4 为例，在训练过程中，每轮迭代计算都涉及前反向传播算法的计算和通信，对超万卡集群的横向扩展（Scale Out）和纵向扩展（Scale Up）网络提出极大挑战。**纵向扩展互联层面，**网络需承载数据并行（DataParallel，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，以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（EqualCost MultiPath）选路不均衡而引发的高速大象流的交换冲突和拥塞。**横向扩展互联层面，**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，横向扩展互联还需要保持高频度、低时延、无阻塞的通信模式。

网络规划需综合考虑 AI 服务器的端口需求和存储需求。以样本面网络为例，其关联计算区和存储区。模型训练时，**一是** AI 计算节点从存储区加载 AI 模型，读取训练数据集。大模型训练过程中训练数据集 batch 读取多以海量小文件为主，以 GPT-3 为例，其训练过程中 70%的小文件在计算节点本地盘命中，仍有 30%需要从存储节点读取。**二是** AI 计算节点通过样本网络将检查点文件和训练模型写入存储区。在 Checkpoint 模型参数保存过程中，为了降低 AI 芯片等待时间，保存时间要求在分钟级，以实现保存模型导致的计算空置率降低至 1%以内。为保障大模型基础设施发挥最大性能，样本网络设计时需综合存储因素。

（四）模型参数急剧增长，开发效率成为新约束

大模型开发全流程技术难度的提升给软件系统在模型训练、模型调优、模型压缩、部署推理等多方面带来了新挑战。大模型训练需处理 TB 级数据，消耗巨量计算资源，调优过程需精细调整超参数，以

平衡模型精度与计算成本。模型压缩需在保持性能的同时，显著降低存储与运行开销。部署时需解决模型在异构硬件上的适配问题，确保高效、稳定地服务于前端应用。推理服务时要求整模型输出结果的实时性与准确性，有效控制推理成本。

大模型训练资源需求普遍较大。大模型参数规模大，与判别式 AI 模型相比，模型训练时计算和存储需求显著增加，依赖分布式技术提升模型训练效率。基于深度学习框架对各类分布式并行策略进行本地化配置，已经成为大模型训练阶段的新挑战。深度学习框架需要有一套简单、灵活、高效且易于使用的框架和工具界面，帮助用户快捷地进行模型训练、调优、配置和管理大规模并行任务。

模型微调、提示工程等增量环节带来开发工具新需求。一方面，模型微调允许模型在特定任务上进行优化，要求开发平台支持高效的模型更新和参数管理。如利用 LoRA 技术可以减少下游任务的训练参数数量，该技术需要开发平台能够灵活地处理模型参数的调整和优化 。**另一方面，**提示工程需输入提示引导模型生成特定输出，要求开发平台能够支持复杂的输入处理和模型交互。同时，提示工程还应用于文本摘要、信息抽取、代码生成等多个领域，要求开发平台提供工具和接口帮助开发者设计和测试各种提示，实现最佳模型性能输出。

超大模型的推理服务需大规模计算资源支持。一方面，超大模型参数给推理端设备带来部署难题。大型语言模型参数量往往达到数百亿甚至千亿级别，如 LLaMA3 70B 参数，需要 136.3G 显存，在推理时至少需要 6 张 V100 GPU 才能有效运行。高昂的计算和存储成本

使得在资源受限的设备上（如移动设备和嵌入式系统）部署这些模型变得极为困难。**另一方面，**超大模型推理响应速度慢，难以适应推理实时性要求。对于严格要求模型响应时间的场景，如语音识别、自动驾驶等，全量大模型推理响应时间慢，难以满足低延时需求。

大模型部署需满足推理侧设备多架构要求。一方面，“英伟达 AI 芯片+Pytorch 框架”体系已成为大模型训练的事实标准和默认规则。英伟达占据全球 AI 芯片市场份额超过 90%，Pytorch 在全球顶会论文中使用占比超过 80%，在 Hugging Face 开源社区中，85%的大模型框架是用 Pytorch 实现的。**另一方面，**国内推理终端硬件种类繁多，大模型在落地过程中面临多样化硬件部署挑战。大模型部署时需适配多种不同硬件平台，包括 CPU、GPU、ASIC 等。此外不同硬件厂商的芯片和软件栈差异巨大，需要将权重文件转换为目标框架和设备可用格式，同时进行一定的优化操作，保证模型统一高效运行满足模型推理需求。

推理任务要求实时性高，推理效率仍需提升。在大模型落地应用场景中，用户期望快速得到响应，若推理速度慢将严重影响用户体验。在金融领域高频交易场景中，交易机会往往在毫秒甚至微秒级的时间内出现和消失，据某知名证券交易所的统计，在交易过程中每秒钟可能有数千笔交易订单涌入，如果大模型推理不能在极短时间内完成对市场数据的分析和决策，将错失大量交易机会，造成巨额经济损失。在自动驾驶领域中，汽车在行驶过程中需要实时处理来自摄像头、雷达等传感器的大量数据，以便及时作出决策，如遇到突发状况时，大

模型推理必须在几十毫秒内完成对数据的分析和决策，否则可能导致严重的交通事故。

（五）基础设施故障率高，运维能力成为新挑战

超万卡集群由数千台智算服务器、交换机、存储设备以及数万根光纤、数万颗光模块构成，大模型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，涉及的软硬件故障模式繁杂，故障管理挑战巨大。

万卡集群训练任务盲启动，失败频发。一方面，千万器件维护管理难度大。单机多卡到万卡集群，系统可靠性“断崖式”降低，器件存在故障隐患，易导致训练任务中断，假设单卡可靠性为 99.99%，“一万+级别”集群，“10 万+级别”光模块，上千万算子，上百套软件适配，集群理论可靠性会降低至 36.7%。**另一方面，大模型基础设施运维需要深度协同 AI 业务。**随着智算集群规模扩大，集群运维管控和集群应用之间的矛盾日益凸显，亟需引入新的运维技术，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建设智能运维系统。运维系统需具备算、存、网、协同管理的端到端系统运维能力，包括计算、存储、网络、开发软件等的管理、控制、分析的全生命周期运维管理能力。

训练过程故障频发，大模型基础设施可用度低。一方面，大模型训练作业中断频繁，业界超万卡集群持续稳定运行时间较短。Meta 的 LLaMA3 在 16,384 个 H100 GPU 集群上进行 54 天预训练，意外中断达 419 次，其中 78%已确认或怀疑是硬件问题导致。Meta 的

OPT-175B 训练，理论上需在 1000 个 A100 上训练 33 天，然而实际训练耗时却达 90 天，其间出现 112 次故障，故障导致手动重启 35 次，自动重启约 70 次。**另一方面，**大模型基础设施故障种类多、复杂系统运维难度大。万亿模型训练过程是多个软硬组件精密配合的过程，故障定界定位复杂，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达数十天。故障发生点涵盖模型训练全流程，算力集群训练前健康检查，训练中的光链路闪断问题检测、深度巡检、跨层跨域快速定界定位等问题均是大模型基础设施运维重点。

三、大模型基础设施关键技术

（一）高效算力管理调度技术

虚拟化、容器化、池化技术成为算力弹性调度基础。一是底层资源虚拟化技术，将物理芯片分割成多个独立的虚拟逻辑单元，为不同的任务匹配合理的虚拟计算资源，有效避免算力浪费。**二是**容器化技术允许在同一操作系统上运行多个隔离的应用实例，极大地减少了资源消耗并提高了资源利用率，尤其适用于快速部署和扩展 AI 应用。**三是**资源池化技术统一管理计算资源，将分散的碎片化资源重新整合，并通过统一接口供上层应用使用，实现了资源的高效调配。

异构并行技术成为算力统一纳管新技术。通常异构并行技术可以适配不同品牌和型号的 AI 加速卡，但异构并行计算实现难度较大。用户在使用资源时需要事先指定使用何种类型的资源，并且只能在该资源池内创建任务，跨架构的资源无法实现并行计算。当前异构并行技术正在加速演进，通过建立“转译”机制等手段，拉齐各异构 AI 芯

片体系的算子、加速指令、通信步调等，使得模型参数与计算框架指令可在异构 AI 芯片之间进行传递并统一执行。

基于预测模型的算力调度体系成为算力调度新选择。一是基于任务资源需求预测的算力调度，充分考虑多个维度资源对用户任务运行效率和基础设施资源利用的影响，构建任务资源需求预测模型，完成自适应资源伸缩调度，有效解决用户资源超额申请问题。**二是**基于业务经济预测模型的算力调度，通过业务经济模型来评估算力资源的产出，使高性能卡向高优先级的作业倾斜，以算力效率和损失决定抢占优先级，在算力分配始终处于高位的同时，实现算力以最经济的方式进行调度，满足收益最大化需求。

（二）高性能大模型存储技术

KV-cache 技术实现长记忆存储，助力大模型推理降本增效。一方面，大模型推理过程中需处理长序列以获取准确的上下文信息、生成高质量输出，模型的计算成本和内存需求通常随序列长度的增加而显著增加，通过 KV-cache 缓存机制，可以有效降低模型长序列推理的内存占用和计算开销，实现有限硬件条件下的高效推理。**另一方面，**基于高性能长记忆存储技术构建的多级 KV-cache 缓存机制，可以保证 KV-cache 具备随时在线和全局池化共享能力，配合以查代算算法，实现从持久化的 KV-cache“长记忆”中调取前期已执行过的计算结果，减少推理过程中的重复计算。根据企业实践，利用该技术可实现推理吞吐提速超 50%，显著降低推理的端到端成本，提升长上下文场景体验。

加速卡直通存储实现数据直达，并行文件系统提高数据供给效率。一方面，加速卡直通存储利用基于总线 P2P 的底层传输协议机制，使数据路径无需再经过 CPU，实现加速卡的 HBM 和存储设备间数据一跳直达，消除 CPU 处理瓶颈，极大地提升了数据从存储到加速卡的传输效率，在检查点状态数据保存、训练数据加载以及 KV-cache 加速等场景发挥重要作用。**另一方面，**具备全局对等共享、单一命名空间和高性能读写能力的并行文件系统，在提高 AI 芯片训练推理效率的同时，实现数据在所有存储节点上均衡分布。同时 TB/s 级带宽和亿级 IOPS 支持能力，可实现万卡集群数据调度简化、供给无瓶颈，保障规模扩展场景下的系统性能。

近数据向量知识库提高大模型检索效率，减少输出幻觉。近数据向量知识库基于快速知识生成、大库容高召回率与多模融合检索关键能力，可实现百亿知识库秒级检索。通过分布式合并建图技术，实现近数建库。根据企业实践，知识生成从月级降至天级，建库时长缩短 5 倍，实现知识实时更新。同时利用存储侧容灾备份特性组合，可提供数据库高可用保障，消除单点故障引发重新建库的巨大开销。

数据编织技术加速跨域数据高效归集，统一数据管理提高全流程效率。一方面，数据编织技术可以借助跨地域全局文件系统，实现全局数据可视可管。跨域集群元数据统一视图、数据入湖后按需流动可以解决多地域、多数据中心的数据资产高效管理和使用难题。**另一方面，**统一数据管理支持 AI 全流程所需的 NAS、对象和并行客户端等协议，呈现跨域跨集群统一命名空间，同时支持无损多协议互通和免

数据拷贝，方便用户管理和访问数据，实现数据放置策略、数据预取能力等全生命周期管理。

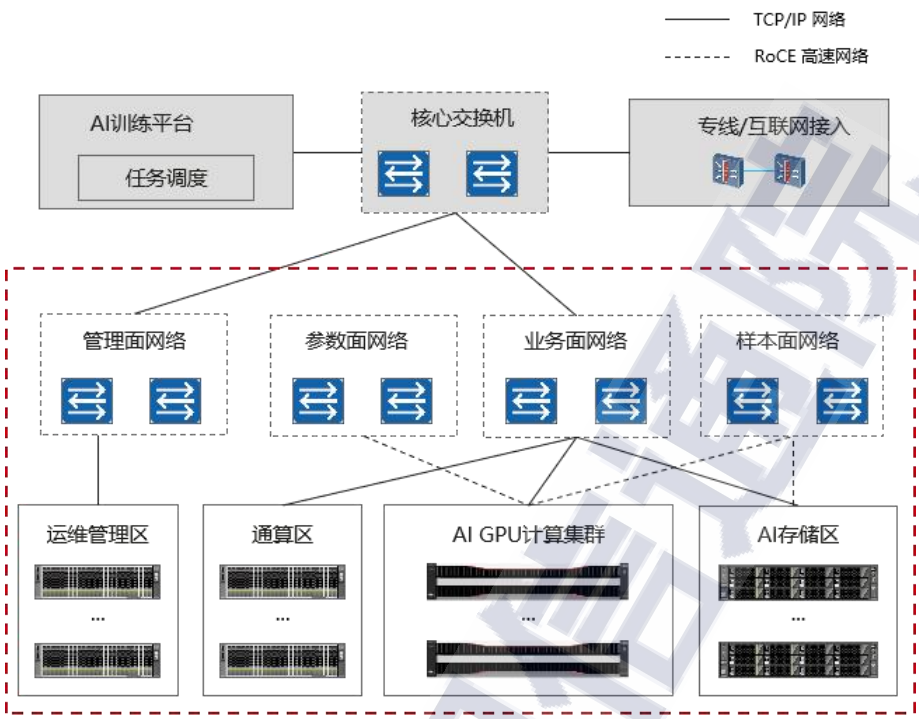
（三）高通量大规模网络技术

高性能互联技术迭代升级，有效提升网络服务稳定性。微软数据显示，万分之一的丢包大致会使算力利用率“变成”九成，千分之一的丢包率会使算力“变成”七成。常见的实现无损以太网络的技术如基于优先级的流量控制等流控技术、显示拥塞控制等拥塞控制技术均依赖于 RDMA 网络技术。当前主要的 RDMA 网络技术，**一是** IB，InfiniBand 协议定义了一套全新的层次架构，从链路层到传输层，不存在 ARP 广播机制等带来的广播风暴问题和组网限制，是目前应用最多的智算网络技术，微软、OpenAI 公司均采用了 IB 网络来进行 AI 智算中心的建设。**二是** RoCE，RoCE 是一种使用以太网承载 RDMA 的网络协议，由 RoCE 规范在以太网上实现了 RDMA 功能，其主要优势在于延迟较低，可提高网络利用率；同时其可避开 TCP 协议并采用硬件卸载，CPU 利用率也较低。

网络负载均衡技术实现最优化流量转发路径，助力解决“算等网”问题。大模型训练时，计算和通信周期性迭代重复，使得在网络侧具有周期性、流少量大、同步突发等流量特点，网络级负载均衡技术可以针对 AI 训练场景下的流量特点，将搜集到的整网信息作为创新算法的输入，从而得到最优的流量转发路径。当前主流解决方案分为逐流和逐包两类。**一是**逐流方案，根据模型计算和通信的依赖关系，对参数传输数据流进行有效调度，合理利用带宽资源，重叠通信和计算，

隐藏通信开销。基于定制化的 xCCL（集合通信库）配合网络路径优化，以及动态多路径负载分担等特性，根据企业实践，在保障网络高可靠性的同时提升网络链路利用率达到 90%以上。**二是**逐包方案，通过自适应路由等技术，报文可以根据网络的负载情况选择最优路径进行转发，从而避免多条流选择同一路径而导致拥塞。企业利用 DPU（数据处理器）解决由于动态选路可能带来的乱序问题，通过乱序重排等机制还原原始流量，可实现整网吞吐达到 90%以上。

参数面、存储面/样本面、业务面、带外管理面网络互联，助力大模型高效训练。大模型训练业务流如图 4 所示，其中涉及计算、网络、存储、AI 开发软件和运维的多系统协调。**训练前，**训练数据集及训练模型需通过存储面网络导入存储系统，AI 开发平台需通过业务面网络和带内管理网络下发训练任务，训练任务镜像、AI 模型、训练数据集需通过样本面网络加载到计算区的 AI 计算节点中。**训练中，**模型参数同步需通过参数面网络，检查点文件需通过样本面网络写入存储系统，若集群故障导致训练中断，则需通过运维系统进行故障排查修复，故障排除后需通过样本网络面从存储系统中加载检查点文件到 AI 计算节点。**训练完成后，**模型通过样本网络写入系统，通过存储网络导出。



来源：昇腾社区

图 4 大模型基础设施网络互联

（四）高效能大模型开发技术

训练加速技术涌现，有效支撑大规模高效构建。一是对计算资源优化，能够通过减少计算和存储需求，提升模型效率，如 PyTorch 支持的混合精度训练、Adafactor、Flash Attention 等技术，目前，谷歌、微软、腾讯、蚂蚁等头部企业广泛采用混合精度训练等技术减少显存占用并提升训练速度。二是计算优化策略，能够有效提升模型的执行效率，如 DeepSpeed 支持的算子融合、梯度积累技术等技术，能够在资源有限的情况下，通过优化计算策略，加速计算过程。三是收敛性优化，能够通过提高模型的收敛速度，提升模型训练效率，提高模型的泛化能力。目前，主流的深度学习训练框架均支持收敛性优化技术，如 DeepSpeed、PyTorch、JAX 等均支持 Adam、Adagrad 等自适应学

习率优化器，能够在训练过程中动态调整学习率，使模型能够更快地收敛。

大模型微调技术出现，模型训练效率进一步提升。为提升大模型在特定场景的适应性，业界推出多种微调技术以提升训练效率。**一是**全量微调技术，微调精度高、泛化能力强，但计算成本较高，一般适用于精度需求较高的复杂任务场景，尤其是需要高精度输出的特定领域任务。**二是**参数高效微调（PEFT）技术，能够显著节省训练时间和计算资源，适用于资源受限或者需要快速部署迭代的场景，已成为产业实践的主流选择。主流的参数微调方法包括低秩适应（LoRA）、前缀调优（Prefix Tuning）、提示调优（Prompt-Tuning）等。当前产业主流的大模型开发平台，如百度千帆大模型平台、阿里百炼大模型平台均支持上述微调技术。

模型压缩技术持续演进，兼顾压缩比例与性能损耗成为首选。大模型通常需经过模型压缩以适应更广泛更多样化的部署环境，以量化、剪枝为代表的压缩技术持续演进，通过低比特量化、稀疏化、模型结构搜索、参数自动寻优等方式实现模型训练中、训练后的低损与高效压缩。如百度的模型自动化压缩工具 ACT（Auto Compression Toolkit）可实现压缩流程自动化，商汤的神经网络量化工具 PPQ 通过图优化等技术实现高效的压缩能力。

大模型推理引擎持续创新，持续提升大模型推理效率。推理引擎针对大模型推理场景的低时延、高吞吐要求，从显存优化、高性能算子、服务调度等多个维度进行优化设计，已成为当前大模型部署推理

的主要工具，如伯克利大学 LMSYS（Language Model Systems）推出的 vLLM、英伟达推出的 TensorRT-LLM、HuggingFace 推出的 TGI、微软 DeepSpeed 推出的 DeepSpeed-MII 等。我国科技企业也纷纷布局该领域，如腾讯推出的一念 LLM 同时支持英伟达 GPU 和华为 NPU 卡，阿里魔搭推出的 DashInfer 聚焦大模型在 CPU 卡上的高效推理，蚂蚁的 GLake 通过对键值对缓存实现透明管理和存算解耦，进一步提升了推理性能和兼容性。

（五）高容错大模型运维技术

训前健康检查，保障作业零隐患运行。一是集群例行检查，在作业启动前触发调用，实现分钟内集群检查。包括芯片健康监测、网络质量检测、环境一致性检测、关联设备告警检查等多项内容。**二是**在集群维护场景下，针对集群环境全量检查，覆盖关键资源的性能测试，如芯片算力测试、带宽测试、HBM 测试、功耗测试、HCCL 带宽测试、leaf 交换机内带宽测试、RoCE 网络连通性检测等。

全栈全路径统一监控分析，资源实时监控。对被管理的计算、存储、网络设备与资源等进行统一日志、指标、跟踪和网络拓扑等运维数据采集，并进行统一管理与检索。提供硬件、网络流量、训练作业和推理服务等多维度的数据可视化与分析。对 GPU 空载率高、GPU 卡异常、掉卡、网络流量异常等进行完善的告警监控。

断点续训，提高模型训练效率。采用基于内存的多级检查点存储方式，实现训练任务分钟级恢复，大幅缩短训练中断时间，提升集群的资源利用率。**首先，**采用异步持久化机制。系统将检查点数据以同

步方式写入内存后，继续执行训练任务。内存中的数据再以异步方式写入持久化存储，从而将训练阻塞时间降低到最小。**其次，**采用内存热加载机制。如果训练作业故障但所在的计算节点状态正常，系统直接从主机内存中加载检查点数据并重启训练作业，减少从持久化存储文件进行读取的 IO 开销。

智能运维，实现故障可预测、可恢复。一是大模型可以深度分析海量运维数据，预测故障发生，自动识别平台中的异常，提前发现平台的潜在风险，并对问题进行根因分析。**二是**大模型可对告警信息进行梳理，定位告警根源，并对告警进行分组，实现告警收敛。**三是**基于大模型 RAG 的运维知识库可以帮助运维人员快速定位问题、提供问题解决的建议，提升运维知识的有效利用。**四是**基于大模型智能体技术，由大模型根据问题和相关运维数据进行决策推理，从而得出问题的解决步骤，随后根据需要选择并调用相应的工具进行问题的自动修复，从而大幅提高运维工作的自动化和智能化程度。

四、高质量大模型基础设施评价指标

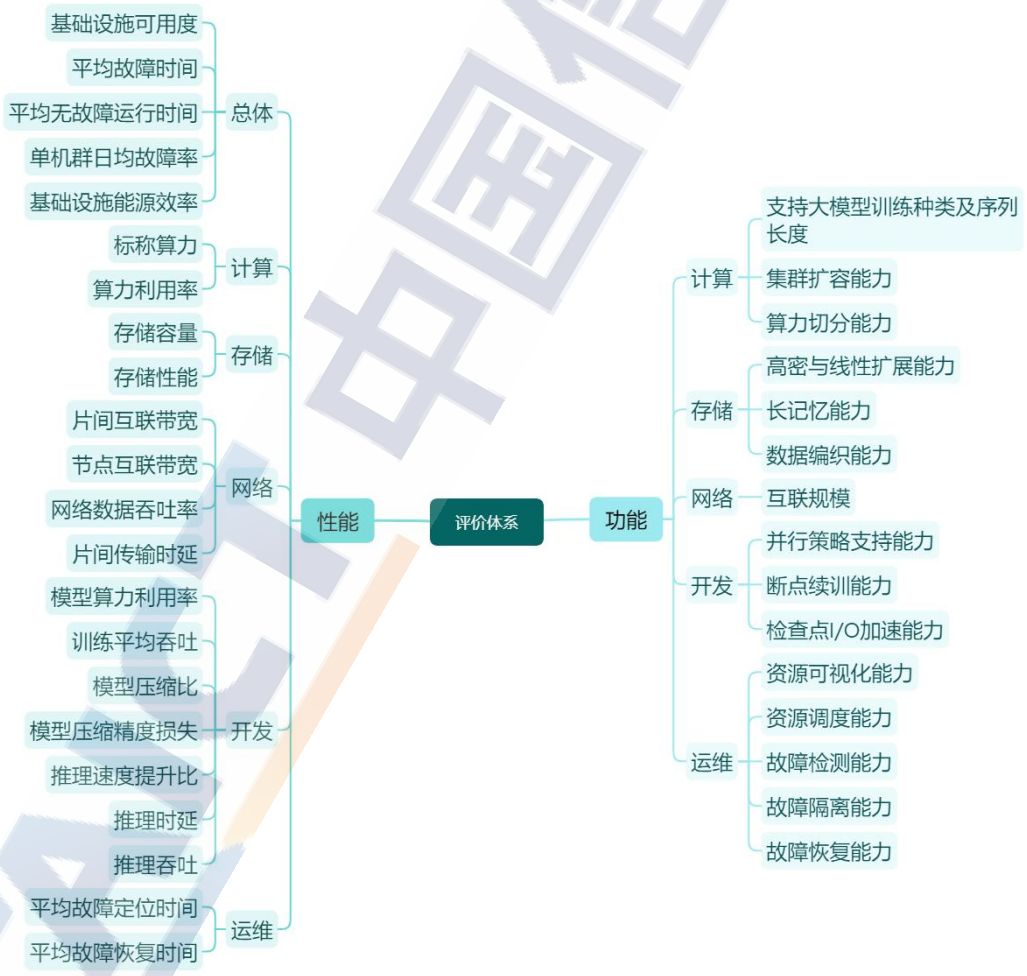
（一）指标体系

大模型基础设施从建好到用好，需要一整套的评价指标。当前大模型基础设施评价指标多集中于计算能力，缺乏对存储、网络、开发软件及运维的系统性评价指标。建立大模型基础设施评价指标，通过明确的指标和评价方法，可以全面了解和评估大模型基础设施的运行情况。

为客观全面地对大模型基础设施进行评价，研究报告从技术能力

和性能层面，聚焦计算、存储、网络、开发工具链及运维等维度建立大模型基础设施评价指标，以技术能力指标和量化的性能指标对大模型基础设施能力进行全面评价，以帮助企业构建适合业务需求的高质量大模型基础设施。

大模型基础设施评价体系需综合考虑大模型的技术能力和性能能力，如图 5 所示。



来源：中国信息通信研究院

图 5 高质量大模型基础设施评价体系

（二）指标定义

技术能力方面如表 1，主要基于大模型全生命周期流程对基础设施的功能性需求建立。计算方面，包括支持大模型训练种类、集群扩展能力、虚拟化能力等。存储方面，包括存储协议等。开发方面，包括并行策略、续训能力等。网络方面，包括芯片互联能力等。运维方面，包括资源可视化、资源调度、故障隔离等。

表 1 大模型基础设施技术能力评价指标

	技术能力
计算	支持平滑扩容集群规模至万卡以上
	训练单卡支持不低于 200TFLOPS（FP16）的算力
	支持训练长序列、多模态大模型
存储	支持容量和性能线性扩展，每 U 不低于 50GB/s 带宽、百万 IOPS 和 500TB 容量
	支持数据编织和加速卡直通存储，跨集群跨地域的数据全局可视可管
	支持长记忆、KV-cache 和近数据向量知识库能力
	支持数据加密，防勒索，6 个 9 以上可靠性
网络	支持万卡以上超大规模智算芯片高效互联
	支持 IB、RoCE 等高速互联技术，支持高吞吐的负载均衡技术
	支持的卡间带宽不低于 200GB/s
开发工具	支持数据并行、模型并行、流水线并行等大模型分布式训练并行能力
	支持断点续训
	支持检查点 checkpoint I/O 加速
运维	支持运维系统集群全局资源可视，故障检测、故障隔离、资源重调度、训练任务恢复全流程自动化

来源：中国信息通信研究院

性能能力方面如表 2，主要基于大模型全生命周期对基础设施的

性能需求建立。系统层面，包括标称算力、集群稳定性、集群可用度等。计算层面，包括计算规模、算力利用率等。存储层面，包括存储容量、存储性能等。网络层面，包括互联带宽、传输时延等。运维层面，包括故障定位时间和故障恢复时间等。能效层面，包括基础设施能源效率等。

表 2 大模型基础设施性能评价指标

一级指标	二级指标	描述
系统	可用度	大模型基础设施集群在一定时间内提供正常服务的时间占总时间的比例，单位%
	平均无故障时间 MTTF	从开始运行到发生首次故障的平均时间，简称 MTTF（Mean Time to Failure），单位小时
	平均无故障运行时间	相邻两次故障之间的平均工作时间，也称为平均故障间隔，简称 MTBF（Mean Time Between Failures），单位小时
	单集群日均故障率	一天内，集群中所有节点发生故障的比率。这个指标通常用来衡量集群的稳定性和可靠性
	基础设施能源效率 PUE	数据中心的总电量÷IT 设备用电量，无量纲
计算	标称算力	指硬件或设备在正常工作状态下的理论计算能力
	硬件算力利用率	模型的实际计算需求与其理论最大计算能力之间的比率，单位%
存储	存储容量	单节点存储容量×节点数，单位 PB
	存储性能	存储系统总带宽（单位 TB/s）和总 I/O 读写次数（单位 IOPS）
网络	芯片片间互联带宽	AI 芯片片间互联带宽，单位 GB/s
	集群节点间	集群节点间互联带宽，单位 GB/s

	互联带宽	
	网络数据吞吐率级	指在单位时间内成功传输的数据量，通常以每秒比特数（bps）、每秒字节数（Bps）或更高的数据单位来表示
开发	模型算力利用率	模型训练过程中实际使用的吞吐量与其理论可用吞吐量之间的比值
	训练平均吞吐	模型在单位时间内能够处理的样本数量
	模型压缩比	压缩后的模型文件大小与原始模型文件大小的比值
	模型压缩精度损失	模型压缩前后，在同一验证集上精度指标之差
	推理时延	从提交请求到收到完成输出的时间
	推理吞吐(每秒查询数)	特定时间内，每秒可成功处理并返回结果的查询请求的平均数量
运维	平均故障定位时间	作业运行时，基础设施集群出现故障到故障首次被发现的平均时间，简称 MTTD（Mean Time To Detect），单位是分钟
	平均故障恢复时间	发生故障后修复所需的平均时间，简称 MTTR（Mean Time To Recovery，MTTR），单位分钟

来源：中国信息通信研究院

五、高质量大模型基础设施典型实践

（一）案例一：Meta 大模型基础设施实践

Meta 认为未来生成式 AI 的训练作业将变得更加多模态化，需要消耗大量文本、图像和视频等数据，基础设施需要满足 EB 级的容量扩展需求，并同时满足高性能和高能效。2024 年 3 月 Meta 披露其最新的两个大模型计算集群技术细节。每个集群均配备了 24576 个

NVIDIA Tensor Core H100 GPU，与既有集群相比在计算、存储、网络、开发软件方面均进行了全面优化，系统架构如图 6 所示。

计算方面，一是改进了任务列表（job scheduler），针对内部作业调度器的网络拓扑感知功能进行多次优化调整，不仅提升了延迟性能，还大幅减少了网络上层的流量负担。**二是**结合 NVIDIA 集体通信库改进，优化网络路由策略，实现了网络资源的最佳利用。任务调度优化实现了任务等待时间减少，集群性能提升 85%。经过一系列优化措施，Meta 大型集群达到了预期的性能水平。

来源：Meta

图 6 Meta AI 集群系统框架图

网络方面，Meta 集群采用两种网络方案。**一是**采用基于 Arista7800 的 RoCE 网络结构解决方案，并配备了 Wedge400 和 Minipack2OCP 机架交换机。**二是**选用 NVIDIA Quantum2 InfiniBand 架构。两种解决方案均支持 400Gbps 端点连接。Meta 通过网络、软件和模型架构协同设计，成功在 RoCE 和 InfiniBand 集群上运行了大

规模的工作负载（包括在 RoCE 集群上持续进行的 Llama3 训练），训练期间未遇到任何网络瓶颈问题。由此可见，RoCE 和 IB 组网的集群均可处理大型生成式 AI 任务。

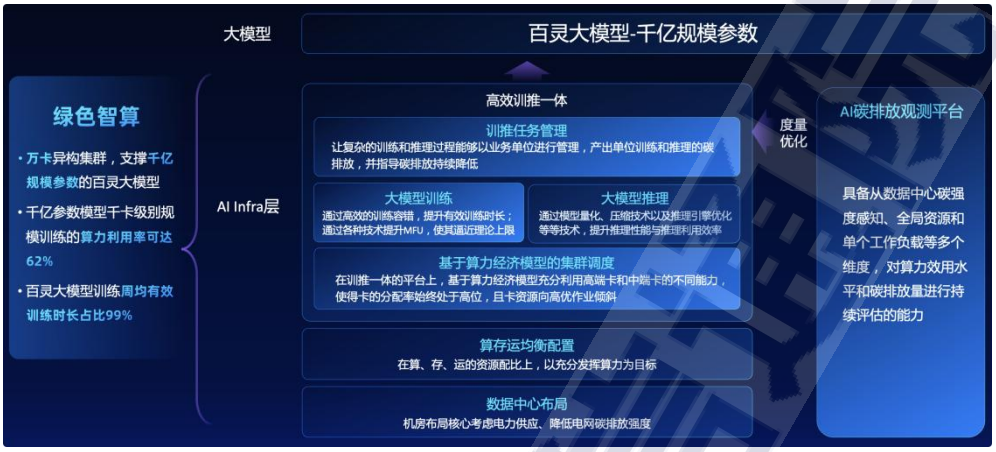
存储方面，一是采用自主开发的基于用户空间的 Linux 文件系统 API，并结合了针对闪存介质优化的 Meta “Tectonic”分布式存储解决方案，实现数千个 GPU 同步保存和加载检查点，同时还实现了灵活、高吞吐量的 EB 级存储。**二是**与 Hammerspace 共同开发部署并行网络文件系统，支持书签 GPU 交互式调试，实现代码更改即时对所有节点生效。**三是**配备最新的高容量 E1.S SSD，实现吞吐能力提升、机架数量减少以及功率效率之间的平衡，同时借助 OCP 服务器的模块化设计，对存储层进行灵活扩展，提高日常维护的容错能力。

软件方面，Meta 利用 MAIProf 识别大模型训练过程中的性能瓶颈，并在 AI 框架层面进行优化。如利用 MAIProf 对模型训练过程中的 Python 函数调用进行全过程跟踪，发现性能异常是因为可配置参数引起，Meta 通过调整可配置参数，在 Pytorch 中使用自动混合精度和多张量优化器进行优化，实现了性能优化。Meta 通过训练框架优化，使得 Pytorch 可支持数万甚至数十万 GPU 并行训练，将优化前的数小时启动时间缩减到几分钟。

（二）案例二：蚂蚁集团大模型基础设施实践

蚂蚁集团构建面向绿色计算的大模型基础设施技术体系，用于支撑蚂蚁百灵大模型（千亿规模参数）的训练和推理。大模型基础设施技术体系包括基于算力经济模型的数据中心布局、算存运均衡配置、

集群调度、大模型训练、大模型推理、训推任务管理等，整体架构如图 7。



来源：蚂蚁集团

图 7 蚂蚁大模型基础设施架构

计算方面，一是全局的编排调度，将数据处理和模型训练协同调度，减少数据搬迁的成本，尽量做到算力和数据的同步，减少算力等待数据的情况。**二是**智算混部，通过共享抢占、动态混部、vGPU 强化等技术，提升算力的利用效率，满足不同业务场景对于智能算力的需求。**三是**建立基于算力经济模型的算力调度体系，通过业务经济模型来评估算力资源的产出，使高性能卡向高优先级的作业倾斜，以算力效率和损失决定抢占优先级，让算力分配始终处于高位的同时，以最经济的方式调度算力，使算力能够发挥出最强效用。

存储方面，利用 KV-cache 技术解决大模型推理显存容量瓶颈与访存密集问题。**一是**提出并实现全新的 KV-cache 显存虚拟－物理映射分配，实现了注意力机制计算过程与显存分配的解耦，使注意力机制核心模块的开发－调试效率从周缩短到小时，最多可减少 71%的显

存占用。**二是**提出并实现对于 KV-cache 的分批显存分配管理（LayerKV），显著降低请求等待时间，使平均 TTFT（Time to First Token）降低 11~69 倍；提出并实现推理显存的动态、弹性分配，使 QPS 提升 1.5~7.9 倍。

软件方面，采用分布式训练加速技术，利用其自研的 PyTorch Native 加速库，集成高性能算子与显存优化策略，适配国产卡，支持完全分布式并行、混合并行等多种分布式策略的灵活切换，支持大规模分布式训练异步更新模式，有效提升训练过程中的算力利用率，测试结果显示，在 Hopper 架构硬件上可平均提速 40%。

运维方面，一是基于 DLRover 实现分布式训练容错。针对大模型训练过程中发现的各种故障进行自动的隔离和容错，无需人工接入，即可实现训练任务的自动恢复。**二是**基于蚂蚁自研的无痛升级技术，实现系统软件升级无需中断训练，保障训练过程的连续性和稳定性。

蚂蚁集团围绕绿色计算理念，构建了面向智能计算的基础设施技术体系，支撑了蚂蚁百灵千亿参数模型的训练和推理，取得了显著成效，其中训练的算力利用率达到了 62%，有效训练时长占比达到了 99%，推理的 TTFT 降低了 69 倍，推理 QPS 提高了 7.9 倍。

（三）案例三：某科技公司大模型基础设施实践

某科技公司为面对业界百模大战，计划快速部署高性能大模型训练平台，以实现大模型的快速上线，抢占市场位置。升级改造前，某公司面临多个集群聚合性能较差，导致模型加载和断点续训检查点读写耗时久，千卡以上集群平均每天故障 1 次，断点恢复时间高达 15

分钟以上，每次经济损失可达几十万元，集群可用度不足 50%。为适应公司大模型发展需求，某公司对原有大模型基础设施进行了系统升级优化。

计算方面，该公司通过自研异构算力调度的大模型训练平台实现万卡集群多机多卡的亲和调度。**一是**通过软硬件协同优化，采用动态张量算子自动融合技术，针对语音大模型的训练进行深度性能优化，使其训练效率达到国际主流芯片的同等水平。**二是**基于虚拟化和池化技术，将异构计算资源汇聚到统一资源池中，进行集中管理与调度，按需灵活分配至不同应用或用户。**三是**通过多硬件联合的量化计算模拟，实现模型在完成单次训练后即可一键部署至不同的硬件平台，大幅提升了模型的部署效率。

存储方面，该公司对存储进行系统级优化。**一是**采用 AI 数据湖存储作为大模型数据底座，通过多套 AI 存储分级建设，实现 PB 级、EB 级容灾扩展能力，同时提升检查点读写能力，断点续训恢复时长从 15min 缩短到 1min。**二是**利用全局文件系统，实现全局统一数据视图和无损协议互通，数据跨域、跨系统数据归集效率提升 3 倍。**三是**通过内生向量知识库，构建全局共享索引，避免跨节点向量查询，实现线性扩容。

运维方面，该公司研发智能拓扑感知、动态负载均衡调度的故障感知与自愈方法，提升大模型训练稳定性。针对大模型训练过程中面临的光模块故障、ECC 报错、流量抖动、掉卡、LOSS 异常等多种问题，通过捕获不同故障码，定义自动故障处理流程，实现了对 200 多

种软硬件故障自动分析和分级处理，80 多种常见故障的自愈时间在 10 分钟以内，4000 余张计算卡任务的连续运行时间超过 20 天。

多系统层面，该公司基于多系统协同优化技术，显著提升了整体性能表现，其中，单机算效提升 50%，通信带宽利用率提升 40%，并行训练算法优化提升 10%，算力集群可用度提升 20%。2024 年初基于此万卡集群发布的商用大模型在语言理解、文本生成、知识问答、逻辑推理、数学能力和多模态能力等多个方面均实现大幅提升，其中语言理解、数学能力超过 GPT-4Turbo，多模态理解达到 GPT-4V 的 90%以上。

六、总结与展望

大模型落地需求推动推理侧大模型基础设施新发展。从应用需求角度看，在众多实际场景中，用户更关心大模型能否快速、准确地给出响应以解决问题。比如在智能客服领域，高效的推理能力可以在短时间内理解用户问题并给出恰当答案，极大提升用户体验和满意度。在医疗诊断中，及时准确的推理能为医生提供可靠的辅助决策，提高诊断效率和精度。**从技术发展趋势看**，随着各行业对实时性要求的不断提高，推理侧的性能优化成为关键。需要更强大的算力支持、更高效的算法以及更快速的数据交互能力。例如，通过优化硬件架构和算法，提高推理速度，满足如金融交易、自动驾驶等对时间极为敏感场景的需求。**从发展需求角度看，**随着大模型的不断发展和普及，对推理侧的可扩展性和适应性也提出了更高要求。不同行业、不同应用场景对大模型的推理需求各不相同，需要推理侧大模型基础设施具备灵

活的配置和定制能力，以满足多样化的应用需求。

绿色低碳将进一步成为大模型基础设施发展重点。从产业需求角度看，企业成本控制驱动绿色低碳技术发展。能源成本在大模型基础设施的运营成本中占比较大。国际能源机构（IEA）的数据显示，2022 年全球数据中心用电量为 2400 亿～3400 亿千瓦时，约占全球最终电力需求 1%～1.3%。根据信通院统计，我国 2022 年数据中心能耗总量 1300 亿千瓦时，同比增长 16%，预计到 2030 年，能耗总量将达到约 3800 亿千瓦时。**从政策推动角度看，**可持续发展策略带来绿色低碳强需求。国家发改委联合网信办、工信部、能源局要求全国新建大型、超大型数据中心平均电能利用效率降到 1.3 以下，国家枢纽节点进一步降到 1.25 以下。

附录 高质量大模型基础设施规划建议

结合国内外业界实践和相关论文，以业界通用的 transformer 类大模型为例，业界大模型基础设施建设主要参考规划。

以常见的 175B 参数量大模型（OpenAI GPT-3，Meta OPT-175B 等）为例，训练数据集 3500B tokens，训练时长要求 50 天。

（一）计算规划

算力规模=8*模型参数量*训练数据量/（训练时长*算力利用率），算力利用率采用较高的 40%，则算力规模=8*175*3500*1000000/(50*24*3600*0.4)=2836PFLOPS

GPU 卡数=8*模型参数量*训练数据量/（训练时长*单卡算力*算力利用率），单卡算力采用常用的 300TFLOPS，算力利用率采用较高的 40%，则：GPU 卡数=8*175*3500*1000000/(50*24*3600*300*0.4)=9453 块。通过提升软件栈、存储和网络能力可以提高算力利用率，进而缩短训练时长。

（二）AI 存储规划

根据英伟达、华为等业界最佳实践，为保障数据高速供给不成为算力的瓶颈，存储性能（存储系统总带宽）（GB/s）≥2×算力规模（PFLOPS，FP16），则存储系统总带宽≥2*2836≈5600GB/s。

存储规模=（NLP 任务数据+ASR 任务数据+CV 任务数据）*（1+元数据占比 5%）/存储水位 70%，单个模型的数据量计算公式如附录表 1。

附录 表 1 AI 存储计算公式

	原始数据	训练数据	CKPT	归档 CKPT
NLP 模型	单个训练数据*100	参数量*20*4	参数量*20*保留天数（10天）*24	参数量*20*存档份数(50 份)
CV 模型	单个训练数据*50	参数量*2*10^5	参数量*20*保留天数（10天）*24	参数量*20*存档份数(50 份)
ASR 模型	单个训练数据*50	参数量*10^5	参数量*20*保留天数（10天）*24	参数量*20*存档份数(50 份)

训练任务分配存储资源时可参考上面的公式，而整体考虑 NLP、CV、ASR、多模态大模型，多租户、多模型并发训练，图片、文本、视频、音频、代码等非结构化数据三个因素，大模型基础设施的存储规模（PB）约为算力规模（PFLOPS，FP16）/40，则总存储规模=2836/40=70PB。

（三）高速网络规划

网络规划主要分为参数面网络、样本面网络、业务面网络和管理面网络，各自独立组网，避免了各个网络之间的竞争和由此产生的拥塞，从而提高系统的可扩展性、安全性和可维护性。其中参数面与样本面网络采用无损以太网，业务面与管理面网络采用普通以太网，各网络平面规划建议如下。

参数面网络：数据并行、流水并行、专家并行均通过参数面通信，

带宽越大越好，采用无损以太网。按照 GPU 卡的卡间支持最大带宽来规划，如 GPU 卡直出接口最大支持 200GE，则 Leaf 交换机选择 200GE 接口以上，Leaf 交换机与 Spine 交换机之间 1:1 无收敛。

样本面网络：用于训练样本的频繁读取、检查点文件的周期性高速保存，以及断点续训场景的检查点文件高速读取，采用高速无损以太网。建议 AI 服务器配置 2*100GE 接口以上，同时存储节点带宽需求高，存储连接 Leaf 交换机采用 2*100GE 以上组网。

业务面网络：主要走控制流、业务部署、外部数据导入等，采用 2*25GE 组网可满足要求。

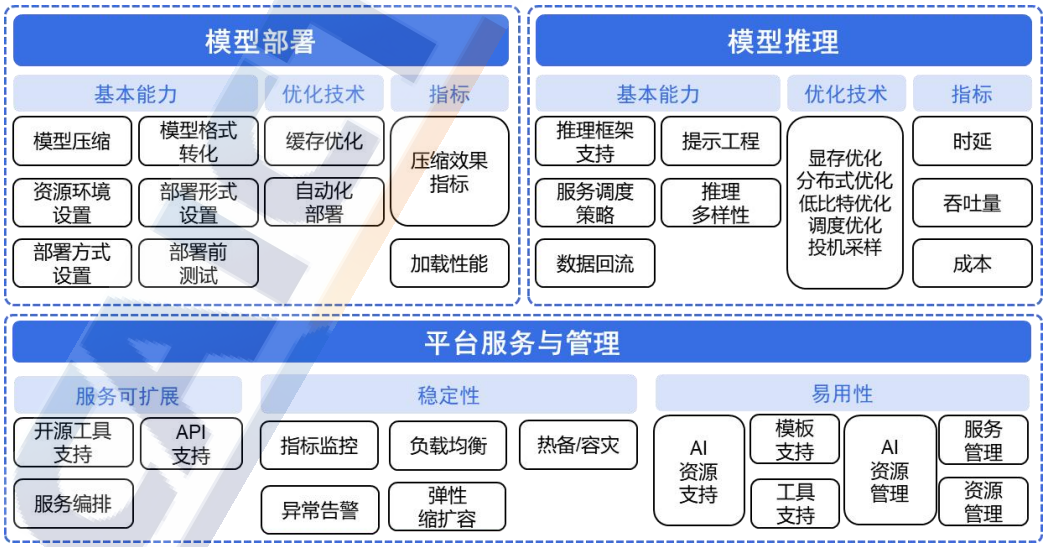
管理面网络：BMC 管理仅设备运行数据，GE 组网满足要求，包含接入服务器、交换机、存储、运维服务器、防火墙等设备。存储带内管理仅配置存储参数，DPC 通信等，GE 组网即可。

（四）开发软件规划

训练微调平台：为应对高参数量、高数据量的模型训练需求，大模型在训练阶段衍生出多种技术能力。大模型训练微调平台应具备分布式训练、端到端断点续训、大模型微调技术、多级缓存和数据加速等多项能力。平台架构参考如附录图 1，能力要求可参考《大模型训练平台技术要求》标准（CCSA 计划号：H-202409119514）。

附录 图 1 大模型基础设施训练微调平台参考功能

为实现大模型推理过程中的低时延、高吞吐、可扩展、易用、低成本目标，大模型推理平台在部署、推理、服务过程中衍生出多种支撑能力。大模型推理平台应具备高效压缩、推理优化支持、服务可扩展能力、服务稳定性、平台易用性等多项能力。平台架构参考如附录图 2，能力要求可参考《大模型推理平台技术要求》标准（CCSA 行标计划编号：2024-1318T-YD）。



附录 图 2 大模型基础设施推理平台参考功能

（五）运维规划

运维平台：大模型基础设施系统复杂、规模大、层次多，大规模、长时间的训练形式和高并发、高吞吐的推理需求对集群稳定性提出高要求，故障可定性成为大模型基础监控、可恢复成为大模型基础设施新挑战。大模型基础设施运维平台需具备资源配置、监控管理、告警管理、运维管理、可视化管理和系统管理，能力要求可参考《大模型计算资源运维平台技术要求》标准（CCSA 计划号：H-202409099502）。

编制说明

本研究报告自 2024 年 8 月启动编制，分为前期研究、框架设计、文稿起草、征求意见和修改完善五个阶段。面向大模型基础设施应用方和技术提供方开展了深度访谈和调研等工作。

本报告由中国信息通信研究院人工智能研究所撰写，撰写过程中得到了人工智能关键技术和应用评测工业和信息化部重点实验室、中国人工智能产业发展联盟、华为技术有限公司、中国移动云能力中心、中国电信数智科技有限公司、联通数字科技有限公司、新华三技术有限公司、蚂蚁科技集团股份有限公司、广东省电信规划设计院有限公司、中兴通讯股份有限公司、软通动力信息技术（集团）股份有限公司的大力支持。

中国信息通信研究院 人工智能研究所

地址：北京市海淀区花园北路 52 号

邮编：100191

电话：010-62301618

传真：010-62301618

网址：www.caict.ac.cn