运营商AI先进存力白皮书

汇聚产业新动能，共创AI新时代

–2023年–

版权声明

编写委员会

顾 问

陈国良、周跃峰

指导委员会

刘景磊、赫罡、魏凯、庞鑫、顾雪军

编 委

陈佳媛、闫晗、童俊杰、李诚、靳震、曹晓峰、董昊、曹峰、丁志彬、何雨今、王振、王旭东、周宇、杨小林、纪烨、韩茂、蒋海林、钟毅、蔡钊、易恩来、蓝文海、谭华、苟欣、沈荣锋、孙睿、江军航、郭洪星、黄维恩、阮政委、孙晓艺、宋天宇、段芳成、钟昭、宋建嘉

主编单位

中国移动通信有限公司研究院、中国联合网络通信集团有限公司、中国人工智能产业发展联盟、华为技术有限公司、中国科学技术大学

序言

人类社会正在跑步进入通用人工智能时代。从AlphaGo 到 ChatGPT，人工智能领域的里程碑事件不断涌现，GPT-4 首次展现极强的语义理解能力、内容生成能力和持续对话能力，这是一个新时代来临的标志。产业界纷纷加强大模型相关领域的研究，并推出一些新产品和新应用，传统信息产业生态正在被重塑。运营商作为 ICT 基础设施建设的主力军，迎来 AI 发展的新机遇。

从对内网络业务角度看，大模型将会加速运营商网络智能化升级。首先，利用人工智能的分析、策略优化与预测等能力来赋能网元、网络等业务系统，有助于提升电信网络的智能规建、智能运维、智能管控能力。其次，通过人工智能设计套餐，将人工智能嵌入用户流量管理中，有助于提升网络运营、市场营销、客户服务的效率。同时，借助大模型还可能对 6G 智简网络以及云网融合的研究提供帮助，促进 6G 技术迅猛发展。

从对外政企业务角度看，大模型也将助力运营商赋能千行百业智能化升级。结合运营商的数据优势、算力优势、行业使能经验优势，运营商将成为数字经济智能化的关键一环。一方面，将运营商的大模型能力外溢至行业客户，面向政务、教育、医疗等推出行业大模型新应用，这也是当前运营商重点发力的方向。另一方面，结合运营商算力、网络等资源优势，为大模型创业者和研发机构提供智算服务，做 AI 淘金时代“卖铲人”。

运营商要抓住大模型的发展机遇，首先需要构建领先的 AI 基础设施，尤其是数据存储能力，也即存力。在大模型场景中，先进数据存力尤为重要。大模型的参数和数据规模都呈指数级增长，对存储的扩展性、稳定性、性能、时延等都提出更高要求。比如一个训练批次前后的数据加载和保存阶段，如果存取性能不足，会导致计算资源（如 CPU、GPU 等）的浪费。存力是算力价值的前提和基础，只有数据存得好、算得快、传得稳，整个 AI 基础设施才能够更好发挥算力的作用。

本白皮书重点研究了运营商如何构建 AI 先进存力，一是分析目前运营商行业 AI 场景的总体发展态势与应用现状，展望了运营商作为国家数字经济发展引擎的重要作用。二是分析支撑大模型应用的 AI 基础设施存在的挑战，分析得出 AI 先进存力是构建领先 AI 基础设施的关键一环。随着 AI 与通信产业的深度融合，AI 大模型不仅会赋能网络的泛在智能能力，还将助力运营商推动千行百业智能化升级。

前行不缀，未来可期。通用人工智能奔涌而来，赋能数字经济的全面智能化升级。运营商引领时代潮流，智算底座将成为千行百业创新变革的重要基石。

陈国良

中国科学院院士

前言

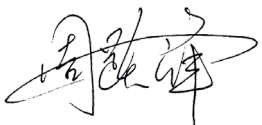
在大模型席卷全球的热潮中，人们已经深刻认识到人工智能作为经济社会发展中一项革命性技术力量，将驱动全球产业实现巨大飞跃甚至跨越式发展，深刻影响未来世界的竞争格局。通信行业作为信息通信基础设施的建设者和运营者，既为 AI 的发展提供基础设施支撑，又将会是 AI 应用落地的先行者。

大模型时代，数据决定 AI 智能的高度。更多的训练数据是 AI 模型迭代升级的前提，更高的数据质量也决定着大模型训练的效果。国内要发展人工智能，并使这一产业得到高速的发展，一定要重视数据和信息的数字化记录。如今，国内建设了大量的数据中心，算力相对较多，但存力较少，很多高价值的信息都没有被记录下来。作为数据的载体，数据存储成为 AI 大模型的关键基础设施。

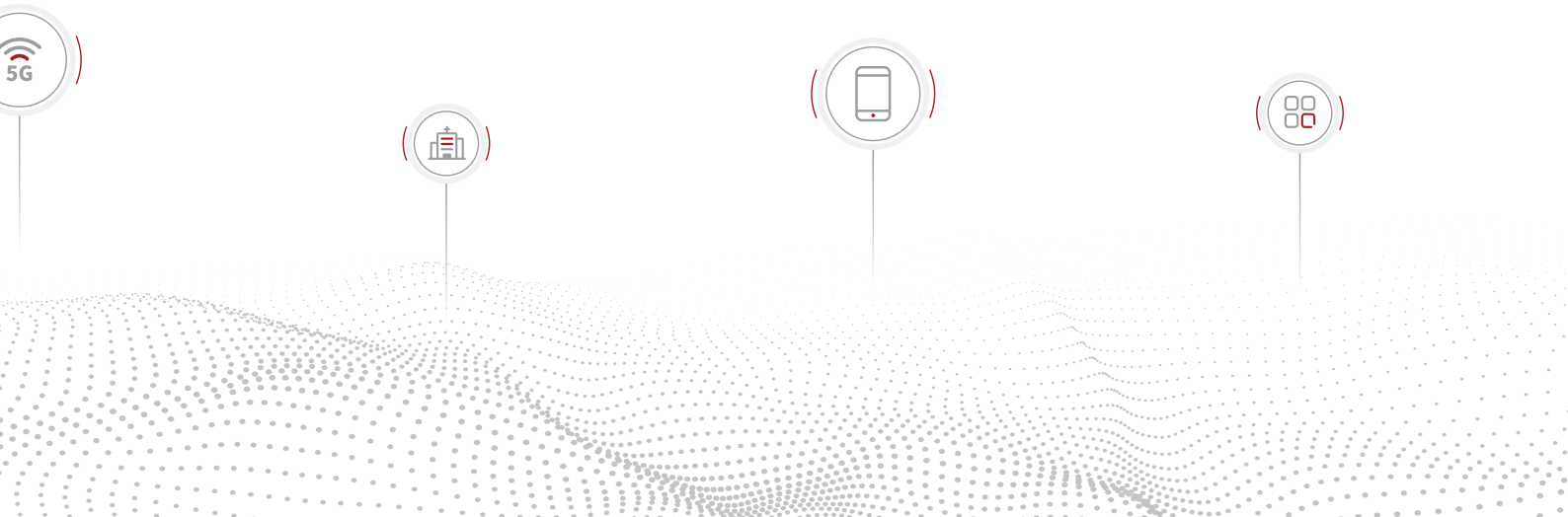
构建 AI 先进存力是构建领先 AI 基础设施的核心条件之一。大模型的持续创新突破，需要从数据的角度对 AI 全流程进行优化。首先，数据归集阶段要高效处理多地域、多分支收集 PB 级多样化的原始数据；其次，在模型训练阶段需要对海量小文件的随机读取以及模型数据集快速保存；最后，在模型推理阶段需要快速检索增量源数据和向量数据。这些挑战都需要创新的 AI 存储解决方案，比如通过智能数据编织，实现跨系统、跨地域的全局统一数据视图和调度；通过近存计算，卸载部分数据预处理能力，减少数据搬迁，缩短数据准备时间；通过全闪存分布式存储，实现存储节点千万级每秒读写次数（IOPS，Input/Output Per Second）和数百 GB/s 带宽，提升训练效率；通过向量存储，实现高性能向量检索能力。

华为公司在数据存储产业上的大规模投入超过十年，提供专业 AI 存储支撑大模型蓬勃发展，助力运营商引领新时代。通过与业界专家、客户和伙伴深入交流，中国移动研究院、中国联通、人工智能产业发展联盟、华为和中国科学技术大学联合编写了本白皮书。白皮书详细阐述了 AI 先进存力对运营商 AI 先进基础设施的支撑作用，建立了一套综合的评估 AI 先进存力的特征体系，可有效牵引对人工智能计算中心的科学规划。我相信这是一次非常有意义的探索，将凝聚更多的行业力量共同推进运营商 AI 产业的发展。

凡人微光，汇聚成炬。华为愿与产业各方更加紧密携手努力，汇聚产业力量，共创运营商 AI 新时代。



周跃峰 博士
华为公司副总裁



图目录

目 录
CONTENTS

1 AI 发展概述

1.1 AI 基本概念

人工智能（AI，Artificial Intelligence）是指通过计算机技术和算法模拟人类智能的一种技术。它可以让计算机像人一样思考和学习，从而实现自主决策的智能化行为。AI 已在计算机视觉、智能语音、自然语言处理等应用领域迅速发展，开始像水、电、煤一样赋能于各个行业。AI 主要分为分析式 AI 和生成式 AI。

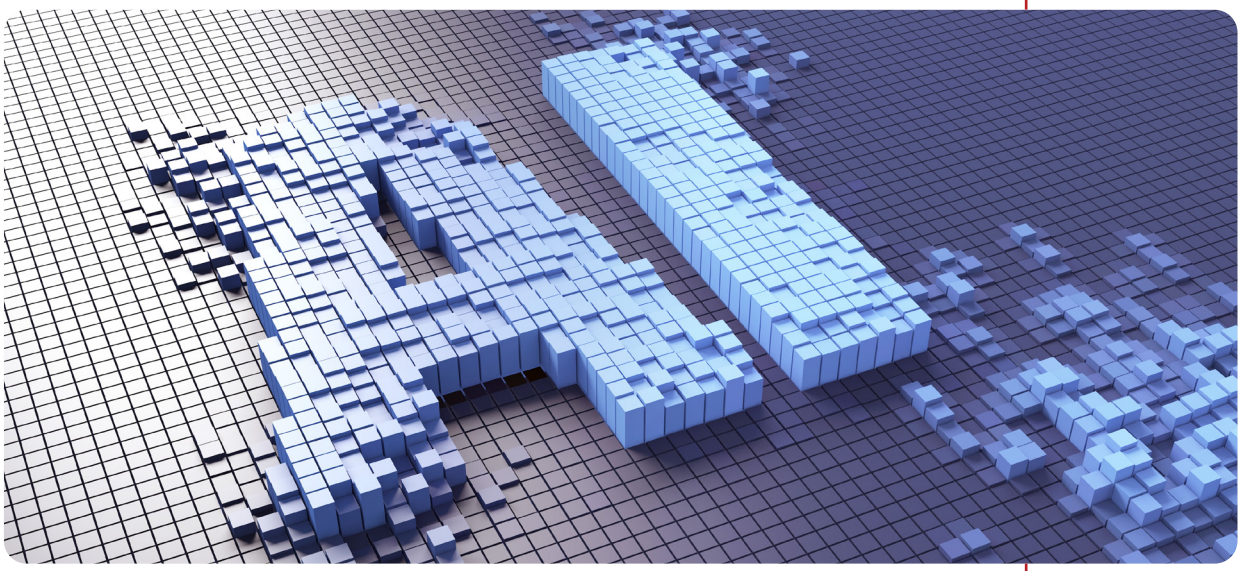
传统分析式 AI 主要用于分析式应用，即对输入内容进行分析和判断，生成输出分析结果，如推荐系统、图像识别、智能语音等。在传统的分析式 AI 时代，只能根据已有的数据进行学习和预测，无法处理新的、未知的情况。

生成式 AI 主要用于内容生成，即使用 AI 生成新内容，如文本、图片、音频、视频等。生成式 AI 在学习归纳已有数据的基础上，学习数据产生的模式，并创造数据中不存在的新样本，实现了从数据分析到内容创作的跨越式发展，打开了 AI 应用新市场，已在文字创作、代码生成、图像生成等多场景实现了应用落地。

1.2 AI 发展阶段

自人工智能科学诞生至今 60 多年的发展历史过程中，人工智能经历了三次发展高潮，分别是 1956 到 1970 年代，1980 到 1990 年代和 2000 年代至今。1959 年 Arthur Samuel 提出了机器学习，推动人工智能进入第一个发展高潮期。此后 70 年代末期出现了专家系统，标志着人工智能从理论研究走向实际应用。80 年代到 90 年代随着美国和日本立项支持人工智能研究，人工智能进入第二个发展高潮期，期间人工智能相关的数学模型取得了一系列重大突破。1997 年，IBM 深蓝战胜了国际象棋世界冠军 Garry Kasparov，是一个里程碑意义的事件。这个时期的 AI 训练数据多为结构化数据，存储方式以本地盘存储形式为主。

当前人工智能处于第三个发展阶段，2006 年加拿大 Hinton 教授提出了深度学习的概念，极大地发展了人工神经网络算法。随后以深度学习、强化学习为代表的算法研究的突破，算法模型持续优化，极大地提升了人工智能应用的准确性。这个时期的 AI 训练数据多为文本、图片、音频等单模态数据，存储方式为本地盘或存储池形式的外置存储。



近年来深度学习发展高潮迭起，Transformer 大模型的问世推动深度学习模型参数量从几万跃升至数千亿甚至更大，模型层数从开始的个位数逐步发展到成百上千，原始数据集也达到 PB 级，为满足大模型对存储的性能和容量需求，外置存储进一步升级为“性能型存储 + 容量型存储”。

大模型实现了更好的泛化能力和更高的性能，正在成为 AI 发展的新范式。一方面，大模型实现了多场景覆盖、精度限制突破、泛化能力增强和研发能力自动化，成为了 AI 走向产业化应用的重要途径。另一方面，大模型也改变了 AI 模型的开发方式，客户无需针对单个场景再开发单独的 AI 模型，而是由 AI 大模型厂商开发基础的 L0 层模型，由行业 IT 供应商开发 L1 层的行业模型，应用场景 IT 供应商提供 L2 层的细分场景模型。

图 1 - 数据 - 模型 - 迭代全生命周期

1.3 我国 AI 大模型发展现状

自2020 年起，我国的大模型在数量和技术水平上都有明显提升，涵盖了智能语音、计算机视觉、自然语言处理等多个领域，并在能源、金融、航天、制造、传媒、城市、社科以及影视等领域应用落地。

从大模型的布局体系来看，科技大厂在算力层、框架层、模型层、应用层进行了四位一体的全面布局。华为、百度均从芯片到应用进行自主研发的全面布局，如华为的“昇腾芯片 + 昇思框架 + 盘古大模型 + 行业应用”、百度的“昆仑芯 + 飞桨框架 + 文心大模型 + 行业应用”。垂直行业企业和科研院所，主要以大模型算法研发和细分领域应用为主。

从大模型参数量看，科技大厂和头部科研机构已上线的大模型参数量普遍在千亿级以上，如阿里通义千问大模型参数在 10 万亿级以上、腾讯混元大模型和华为盘古大模型参数量均在万亿级以上、百度文心一言大模型参数量在 2 千亿级以上，上海人工智能实验室书生浦语大模型参数量在千亿级别。

从大模型业界评价看，国内大模型处于百花齐放状态。综合实力方面，头部科技大厂如华为、百度、阿里、腾讯等在大模型研发投入、技术能力和人才团队方面均有较强实力。商用推进方面，头部科技大厂均依托现有业务领域进行大模型应用推广，更容易形成大模型的规模化应用。

2 运营商是助推 AI 强劲发展主力军

2.1 运营商高度重视 AI 发展

电信运营商作为 ICT 基础设施的建设者和运营者，拥有全国跨域网络互通、云网融合统一纳管的资源优势，也担当着全社会数智转型主力军的重任，因此在人工智能产业飞速发展的过程中，也在抓住新机遇加快构建新一代 AI 基础设施，提供澎湃算力、先进存力、品质运力，赋能千行百业。

1. 中国移动 AI 战略布局

中国移动在 2013 年开始人工智能领域战略布局，现已形成从算法、平台、到规模化应用的产业级智能化服务能力。其九天平台已具备在计算机视觉、自然语言处理、智能语音、网络智能化等多领域的 AI 能力，跻身央企“AI 国家队”。

中国移动已发布九天海算政务大模型和九天客服大模型。依托九天海算政务大模型，政务服务系统将具备强大的政务事项理解能力、多维度的信息关联能力、面向复杂事项和复杂流程的多元交互能力。九天客服大模型既可根据用户提供的自然语言描述，解析用户问题并提供答案，又可以与人工客服协作，为人工客服提示回复建议，形成“大模型—人工坐席—用户”的三方沟通场，极大提升人工客服的工作效率。

AI 技术也已经在中国移动的多个业务领域实现规模化应用，助力管理、服务等多方面能力提升。例如智能客服月交互量从 5000 万提升至 2.1 亿，准确率达 92%；反诈骗系统月度拦截电话量超过 1400 万，准确率高达 98%；声纹识别防欺诈防骚扰电话准确率已达 98%。

中国移动正在围绕智慧网络构建国家新一代人工智能开放平台，加速“通信网络 + 人工智能”的关键技术突破。一是围绕网络规划和运维业务，建设智能化仿真实验环境。利用人工智能技术，提供覆盖优化、多目标天线优化、大话务量业务保障、无线 CSI 压缩及反馈等业务。二是面向社会开放多场景 AI 基础设施，如对 ICT 企业、高校、科研机构、行业组织等提供算力、数据、算法、平台等资源。

同时，中国移动在 2023 年 8 月发布的《中国移动 NICC 新型智算中心技术体系白皮书》中对“新存储 - 挖掘数据价值”做了详细阐述，通过计算与存储的交互过程总结出智算场景存储面临的性能，容量和调度关键挑战，最后提出多协议融合存储贯通异构数据，全局统一存储打破单体局限和基于计算总线构建统一内存池等解决方案。

2. 中国电信 AI 战略布局

中国电信在 2019 年到 2020 年期间，确定了云网融合人工智能发展战略，先后发布了《中国电信人工智能发展白皮书》、《云网融合 2030 技术白皮书》，同时围绕业务中台、数据中台、安全中台、原子能力平台和云网技术底座提出“三中台一平台一底座”的数字化平台顶层架构，明确了云网蓝图。

基于云网融合优势以及天翼云多年的技术沉淀，中国电信推出智能计算平台“云骁”，提供智算、超算、通算多样化算力服务。依托天翼分布式架构云底座，“云骁”可提供软硬一体的解决方案，实现高阶算力供给、资源高效利用，助力行业数字化转型，降低企业创新成本。

中国电信于近期推出星河通用视觉大模型 2.0，旨在为状态检测、动作事件、工业生产等场景提供服务。星河大模型参数量已从 10 亿提升至 100 亿，并融入图像、视频、语义多源信息，其语义理解能力、视觉感知能力、精细分割和空间交互关系能力均得到进一步提升。

天翼云智能计算平台还为客户提供大模型训练和微调服务。通过“云骁”平台提供分布式训练一站式解决方案，进一步缩短模型交付周期、提升 AI 训练开发效率。

此外，针对技术合作伙伴，中国电信提出了“云创计划”，云存储是云创计划的 5 个领域之一。云存储重点解决多场景存储问题，聚焦数据存储搭建、融合存储合作。

3. 中国联通 AI 战略布局

中国联通的 AI 应用战略包括两部分。一是对内提供智能化运营，如 5G+AI 智能运营平台，利用 AI 提供网络故障定界问题能力。二是对外提供一站式创造服务，如一站式 AIGC（Artificial Intelligence Generated Content，人工智能生成内容）创造工厂。

中国联通已经发布鸿湖图文大模型 1.0，其具备以文生图、以图生图、视频剪辑等功能。随着移动互联网的快速发展，用户对于个性化、原创性的内容需求也越来越高，传统的图像、视频生成方式无法满足用户的需求，鸿湖图文大模型的推出填补了这一空白。通过该模型，运营商可以为用户提供丰富、有趣的图文内容，进一步提升其增值业务的竞争力和用户体验。

鸿湖图文大模型的应用潜力巨大，可被广泛应用于媒体、广告、娱乐等多领域。媒体领域，鸿湖图文大模型可以帮助媒体机构高效、快速地生成新闻稿件配图；广告领域，鸿湖图文大模型可为广告公司提供广告创意和广告图像；娱乐领域，鸿湖图文大模型可以为用户创造丰富多彩的虚拟世界。

联通还在全力打造 uniVerse 元宇宙平台，推出一站式 AIGC 创作工厂——联通元宇宙 AIGC 平台。该平台是联通面向 AI 商业落地布局的重要一环，包括 AI 跨模态检索、AI 图片驱动、AI 音乐、AI 主播、AI 绘画、AIGC 3D 数字人等多种功能。

此外，联通云 7.0 面向 HPC/AI 场景推出文件存储系统，目标是做 AI 时代的良田沃土。相对传统存储显著进步的地方有三点，一是更高的性能，包括高吞吐量以及部分计算场景下要求非常低的时延，以减少计算集群等待时间，让平台持续高效的运转；二是可扩展，AI 时代下，存储的性能可扩展、容量可扩展成为承接巨量数据、高增速、高性能计算要求下的硬性要求；三是多接口，包括 POSIX、S3、ISCSI 等协议接口。

2.2 运营商应用 AI 的主要方向

1. 对内融入现有业务，提升业务效率

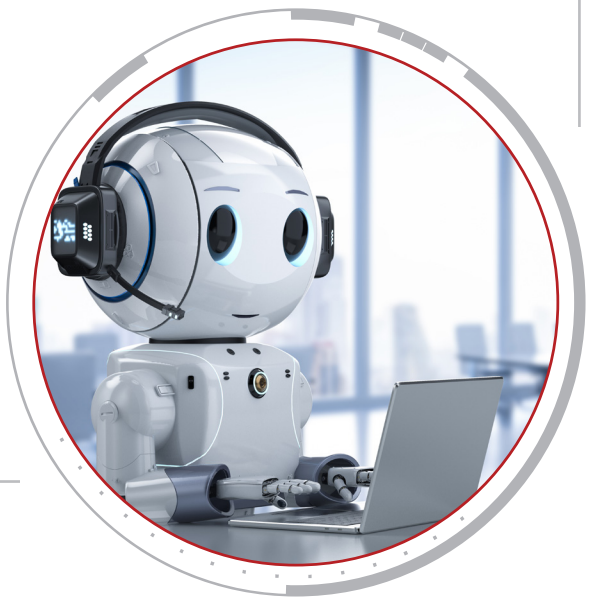
AI 应用与运营商现有业务结合，实现业务效率提升。通信网络侧，AI 技术可以快速定位网络故障、简化网络优化流程，让运营商网络更加安全、稳定、可靠。客户服务侧，AI 技术可以帮助运营商更好地满足客户服务需求，增加客户参与度，提升用户体验。依托人工智能的语音识别、自然语言处理、人脸识别、知识工程等技术，运营商可以让 AI 技术与现有业务结合，大幅提升运营、运维效率，改善用户体验。

网络优化方面

AI 能够在移动网络和固定网络“规 - 建 - 维 - 优”的各个环节得到应用。网络智能配置方面，人工智能技术结合网络历史数据，将专家经验数字化，通过对网络性能的预测和自动化操作配置，有望实现移动站点智能规划、基站业务快速开通、智能路径规划和光传送网自动化部署等应用。网络智能运维方面，人工智能技术可实现物联网端到端质差识别定位、无线网络异常小区发现、IPRAN 故障分析定位等应用，可有效减轻运维人员负荷、提升运维故障处理效率。网络智能管控方面，人工智能技术可基于网络历史数据实现多种应用，如智能频谱管理、智能切片管理、智能负载均衡、智能缓存管理、智能路由、自适应传输功率控制与传输质量管理等。网络智能优化方面，人工智能技术可实现网络的主动优化和全局优化，包括移动性管理增强、智能基站节能、无线网策略参数智能优化、智能路径优化等。

客户服务方面

AI 技术可以降低人工客服中心的负载，减少客服中心的成本，提高客户满意度。同时，AI 能够在智能语音助手、坐席助理、智能推荐、自助服务、社交媒体管理、个性化服务等多个场景提高客户服务的质量和效率，满足客户日益增长的个性化需求，帮助企业更好地服务客户，提高竞争力和盈利能力。以 ChatGPT 为代表的大模型技术的出现，将会加快智能客服的发展，在语义理解、情感识别、知识搜索定位、客户体验等方面提升效率和体现。



2. 对外赋能产学研用，推动智能升级

AI 大模型作为数据、算力、算法三位一体的产物，对 AI 基础设施的需求高、投入大。以 OpenAI 为例，根据公开资料，ChatGPT 初期估计投入高达 8 亿美元，GPT-3 的训练总成本也高达千万美元，一般企业很难承担如此高昂的基础设施成本。运营商可以将自己的 AI 基础设施（AI IaaS）、平台能力（AI PaaS）、AI 模型（AI MaaS）以服务的方式租赁给 AI 创业者和研发机构，帮助企业降低 AI 业务开发的难度和成本。运营商利用自身的网络、用户、平台和数据等优势，能够更好的面向企业、政府发挥作用，打开新的市场空间。

中国电信启动了大模型生态合作联盟。该联盟将推出数据合作计划、亿元算力扶持计划、千万创新激励计划、品牌支持计划、渠道支持计划、资本赋能计划六大合作伙伴政策。同时，中国电信表示将依托云网融合优势，打造四级智能算力体系，围绕行业应用需求，联合产业链上下游生态合作伙伴，为千行百业量身打造定制化的行业应用大模型，促进各行业数字化、智能化转型升级。

中国移动的九天海算政务大模型首创“政务大模型 - 信息场 - 应用”端到端政务服务体系，一网通办的服务理念将被更加安全、高效地体现在群众的办事体验中。对于政府工作人员，通过大模型和信息场的联动，快速获取直观的数据分析结果，满足工作人员动态管理、公文写作等需求，实现跨层级、跨地域、跨业务、跨系统、跨部门的高效政务治理体系。中国移动将构建九天大模型全方位合作生态，联合顶尖高校、科研机构加速大模型关键技术创新，汇聚优秀基础软硬件伙伴，打造先进人工智能算力基础设施，携手业界龙头共建大模型，共助千行百业智能化应用创新。

中国联通的鸿湖图文大模型可以实现文本生成图像、视频剪辑和图像生成图像等功能。中国联通还携手华山医院、上海超算中心等单位发布了“Uni-talk”医疗算网大模型。该大模型是一款为医疗行业定制的大模型产品，功能类似 ChatGPT，不过更加聚焦于医疗领域的应用。华山医院会将“Uni-talk”应用于专业医学文献检索，辅助诊断等场景。

“算力、数据、算法”三位一体的驱动下，运营商有望迎来 AI 与数字经济带来的第二增长曲线，AI 发展带来的流量增加，将会直接带动运营流量收入上行。运营商作为数字时代的中坚力量，其价值也将进一步凸显。

3 运营商 AI 存力挑战

3.1 AI 存力是运营商发力大模型的基础

当前我国三大运营商均已发布各自的 AI 战略、AI 平台、AI 大模型，甚至是行业大模型。运营商在全面拥抱 AI 新机遇时，会充分发挥其既有的数据优势、资源优势、行业使能经验优势，全力打造领先的 AI 基础设施，依托算网融合的能力，让 AI 服务无所不达。面向大模型应用的 AI 基础设施除了对算力有极高的要求外，对网络传输能力和数据存储的能力也提出了更高的要求。中国工程院院士倪光南指出“存力、算力、运力缺一不可，只有三者平衡配置、均衡发展，才能充分发挥算力的作用”。

从运营商对内融入现有业务降本增效来看，需要存力系统具备数据统一调度能力。运营商首先要汇聚各地域的网络数据，然后基于最新的业务系统运营情况及时更新数据。这些数据可能来自不同地域的业务系统、不同厂商的公有云或私有云、不同的合作单位或生态伙伴。这就需要构建支持全局统一数据视图的存力系统，以解决跨厂商、跨地域、跨云的数据统一调度问题，为大模型注入源源不断的数据“燃料”。

从运营商对外赋能千行百业智能化升级来看，需要存力系统满足低时延、大带宽。运营商需要高效训练出各类行业大模型并提供敏捷、精准的推理服务。大模型的训练周期长、训练数据量大、训练流程业务模型差异大，需要运营商具备同时满足极低时延、超大带宽、混合负载自适应均衡的存力系统，减少计算对数据读写的等待，为行业大模型提供动力澎湃的数据“引擎”。

从运营商提供大模型服务来看，需要存力系统满足数据全生命周期的高安全、高可靠。大模型在训练和推理过程中产生的关键节点数据和最终的模型文件数据都是企业的核心资产。一方面需要保障 AI 服务的高可用性，避免被设备故障或自然灾害中断，另一方面还需要保障核心数据资产的安全，避免被病毒勒索等人为攻击。为此，运营商需要构建端到端基于数据全生命周期的可靠存力系统，为大模型提供坚实可信的数据“护盾”。

综上所述，存力为大模型提供海量数据存储支撑和高效训练推理支撑，同时又为安全可靠的大模型服务保驾护航，是运营商发展好大模型最重要的基石之一。

3.2 运营商 AI 存力面临的主要挑战

运营商丰富的 AI 应用对传统 IT 基础设施带来了全面的挑战，运营商依据 AI 业务流（数据获取、数据预处理、模型训练、推理应用）独立建设存力设施后，设备多、版本多、冗余数据多、数据管理复杂等问题不断累积，进而出现了数据容量、数据传输、数据管理、数据安全、数据节能等维度的 AI 存力难题。

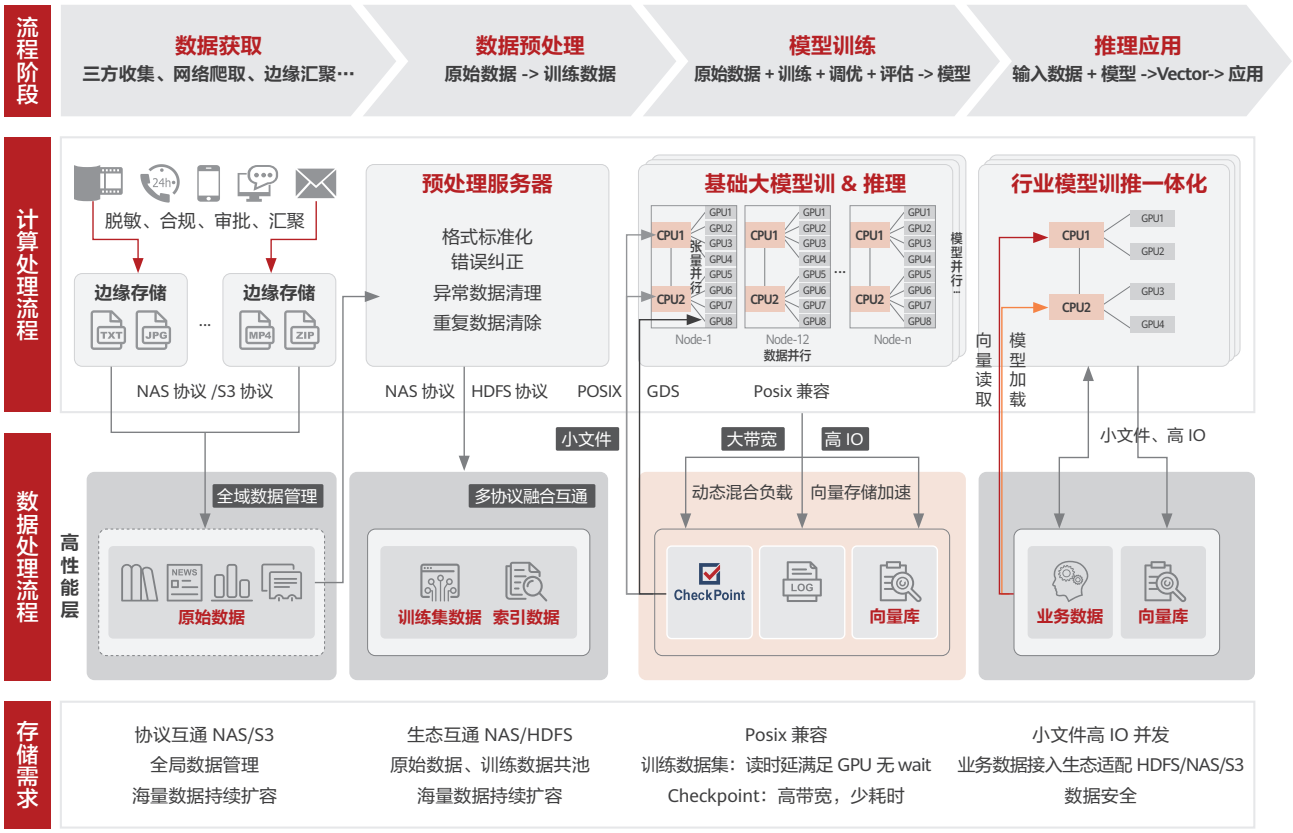
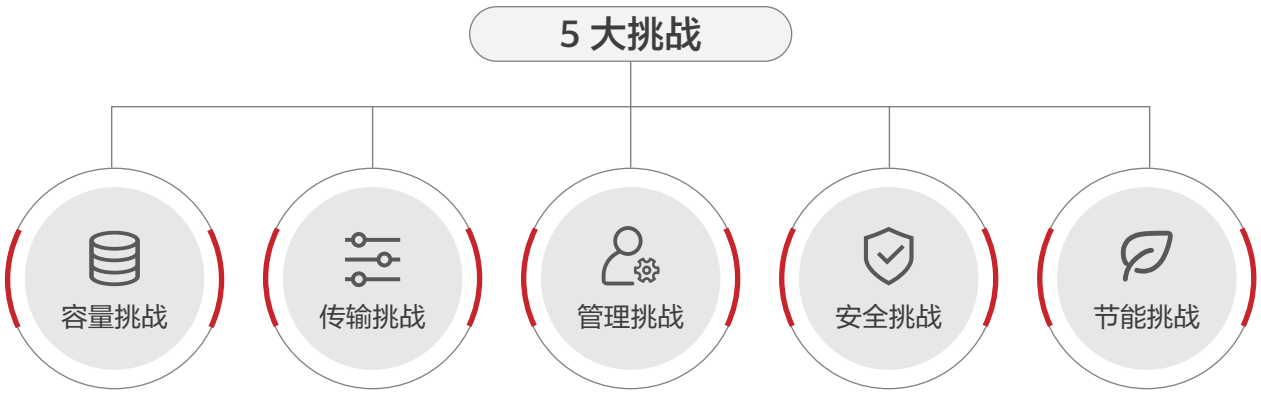


图 2 - AI 大模型全流程存储需求分析



1. 容量挑战

运营商传统的 IT 基础设施主要应用于超算、大数据等领域，无法满足 AI 大模型对存力平台的诉求。容量层面主要面临以下三方面的挑战，一是灵活性不足，随着大模型的发展，模型进入万亿级参数、PB 级存储时代，这使得传统的单机模式和服务器盘集群建设难以满足千亿级文件系统管理、PB 级存储持续扩容以及数据响应性能要求。二是开放性不足，多模态大模型需要 AI 存力设施支持多样性的数据存储需求，传统的建设方式需要为每一种新的数据类型配备对应的存储设备，这限制了系统的开放性和灵活性。三是建设难度大，AI 业务流在数据获取、数据预处理、模型训练和推理应用的各个阶段对数据存储的诉求差异极大，传统的独立业务域存储建设模式难度极大。

因此，新型的 AI 存力一方面需要同时支持 NAS/S3/HDFS/POSIX 等传统数据服务接口，另一方面需要持续创新，支持训推阶段向量信息的新型存储格式。数据格式多协议互通互享，可以有效降低数据存储空间，并提高数据管理效率和向量数据访问性能。

2. 传输挑战

（1）数据跨域调度

运营商在发展 AI 大模型业务时，需要获取来自不同机构的多样化数据，包括互联网、企业内部、合作机构和省分公司等。为了最大化利用存储能力，需要实现高吞吐和大容量数据传输。然而，传统的建设模式选择不同的计算资源、网络资源和存储资源构建基础设施底座，多样性的设备带来了业务难迁移等问题，使数据难以流动。因此，新型的 AI 存力一方面需要建设统一的数据湖存储实现大容量的数据传输。

（2）数据高效流动

大模型训练任务对内存和显存带来较大挑战，数据需要在计算、Cache、高带宽内存（HBM，High Bandwidth Memory）、DDR 内存设备之间频繁移动，缺乏统一内存空间的寻址会导致编程模型变得复杂，也会限制设备之间的协作，增加了开发难度和错误率。同时在 DDR 内存和 HBM 之间数据需要多次转换，异构设备既无法直接共享数据，也无法充分发挥各自的优势，这些因素都限制了系统整体性能的提升。因此，需要引入统一的内存引用方式和服务调用接口总线技术，如灵衢总线（UB ，Unified BUS）或 CXL（Compute Express Link）等技术。这种总线技术提供了基于内存语义的数据中心资源池化和高效共享机制，允许程序地址的直接引用，并支持分布式执行的远程功能调用，从而满足了包括 AI 大模型、大数据分析和云超算等在内的多种紧耦合、大规模、高性能计算需求，有助于数据中心高效率编程，从而极大地提高了数据中心的性能和效率。

3. 管理挑战

大模型从单模态走向多模态，多样化的数据类型给数据存储管理带来了巨大的挑战。一是大模型训练需要复杂的文件读写，数据存储系统不仅需要支持千万级 IOPS 和数百 GB/s 的带宽诉求，而且需要在技术和管理方面进行不断的改进和创新。二是大模型训练面临计算处理能力瓶颈。大模型训练时需要在 CPU 上执行复杂多阶段的数据预处理流程，包括提取、转换、加载等，如何通过存储系统管理降低 CPU 的数据预处理负担是新的挑战。

因此，新型 AI 存力需要从数据全生命周期管理的角度解决上述问题。一是基于全域数据管理发现无用数据、冗余数据、热温冷数据等；二是提供数据分布视图并指导用户进行数据存储的重新规划，减少存储开销，同时支持数据和模型云边调度及推送能力；三是构建全局虚拟数据总线，为 AI 平台提供全局数据空间，以及安全、高效、易用的数据存力网络。

4. 安全挑战

运营商的人工智能计算中心是国家的重要基础设施，是推动科研创新和工业发展的关键动力。AI 存力平台是人工智能计算中心的重要组成部分，在安全保障体系建设过程中，需要满足供应安全和自主可控，包括国产控制 CPU、系统管理芯片、接口卡处理芯片、固态硬盘控制芯片，以及自主可控的数据存储介质等，从根本上保障 AI 先进存力平台的供应链安全。

AI 存力平台汇聚了海量数据和高价值大模型文件，需要维护数据的机密性、完整性和可用性。一是在应用层提供安全的开发环境，如账号安全性、数据安全性、权限安全性、编码安全性等能力。二是在系统层对使用的操作系统、数据库和中间件容器等进行安全保护。三是在网络层对网络设备以及通信提供保护。四是在管理层对系统维护、运作活动进行监管和保障，确保系统安全的连续性。

通过各个层面的分工配合，可以为业务系统提供安全、可靠、稳定的服务，并为客户提供资产安全保障。然而，对于数据存储系统来说，还需要进一步做好数据层面的防护。数据安全的威胁主要集中在数据破坏、数据泄露和数据勒索等方面。为了防止数据被破坏、被窃取和被勒索，需要有效地识别攻击，并在受灾后恢复数据。

5. 节能挑战

运营商是实现“双碳”目标的重要力量，电信业务系统的数据迁移会产生能源消耗，在业务全周期内减少数据迁移次数，能够大幅优化能耗开销。然而，当前 AI 业务系统内部的多类数据迁移操作相互独立，缺乏对数据系统级和 AI 作业生命周期级别的数据排布规划，导致数据迁移的代价较大，严重影响了系统效率，消耗了大量的存储计算资源。因此，如何在系统架构上减少低效数据处理和无效数据迁移所带来的巨大能耗开销，是当前面临的关键挑战。

在碳中和目标下，运营商节能减排和绿色转型势在必行。面向未来的存力平台需要结合 AI 业务特征进行能耗优化。一是宏观架构上存算协同，计算、存储资源独立部署，通过高通量全局虚拟总线互联，虚拟总线统一内存语义访问数据，实现计算、存储资源灵活调度和利用率最大化。二是微观上存算一体，减少数据非必要迁移，在数据产生的边缘、数据流动的网络、数据存储的系统中布置专用数据处理算力，根据业务支持算子下推，将其卸载至 DPU、内存、盘控制器，提升数据处理效率。

4 AI 先进存力发展趋势

4.1 AI 先进存力内涵

随着大模型时代到来，训练数据量与模型参数呈指数级增加，更复杂的 AI 业务流对 AI 基础设施的存力提出了更高要求。数据存力朝着大容量、高性能、安全可靠和绿色低碳的方向不断发展。业界认为 AI 先进存力是一种以“大容量、高 I/O 性能、高带宽性能”为基础，以“开放生态、高效架构、先进介质”为支持，具备“绿色低碳、安全可靠”等特征的存储能力，它能够实现存储系统的敏捷扩展，支持数据服务开放共享和数据业务高可用，保障系统可持续发展和数据隐私安全。

AI 先进存力可以作为人工智能计算中心存储能力建设的参考，其至少应该具备以下能力，一是支持大容量的敏捷扩展，包括 PB 级容量的敏捷扩展和大小 I/O 混合负载自适应能力；二是支持数据开放共享，满足 AI 业务不同阶段数据管理、数据流动的需求，加强数据服务的开放性和互联互享能力；三是支持数据的高可用，满足更强的数据服务可用性要求，保障全生命周期业务；四是支持数据的隐私安全，满足数据资产和数据隐私的平衡要求，加强全生命周期隐私管理；五是支持存储的高效节能，通过提升数据处理效率以达成节能。

4.2 AI 先进存力关键技术

以数据为中心，统一存力基座，融合多元异构算力的新 AI 技术架构，已逐步成为人工智能计算中心的主流架构。多种异构算力紧密围绕在统一的数据底座，改变了过去“数据跟着算力跑”的算力烟囱工作模式，朝着“算力围着数据转”的新模式演进。作为数据载体，数据存储已成为构建 AI 大模型的关键基础设施之一。为了构建先进的 AI 存力，需要从存储介质、架构、设计、安全和低碳等方面发力。

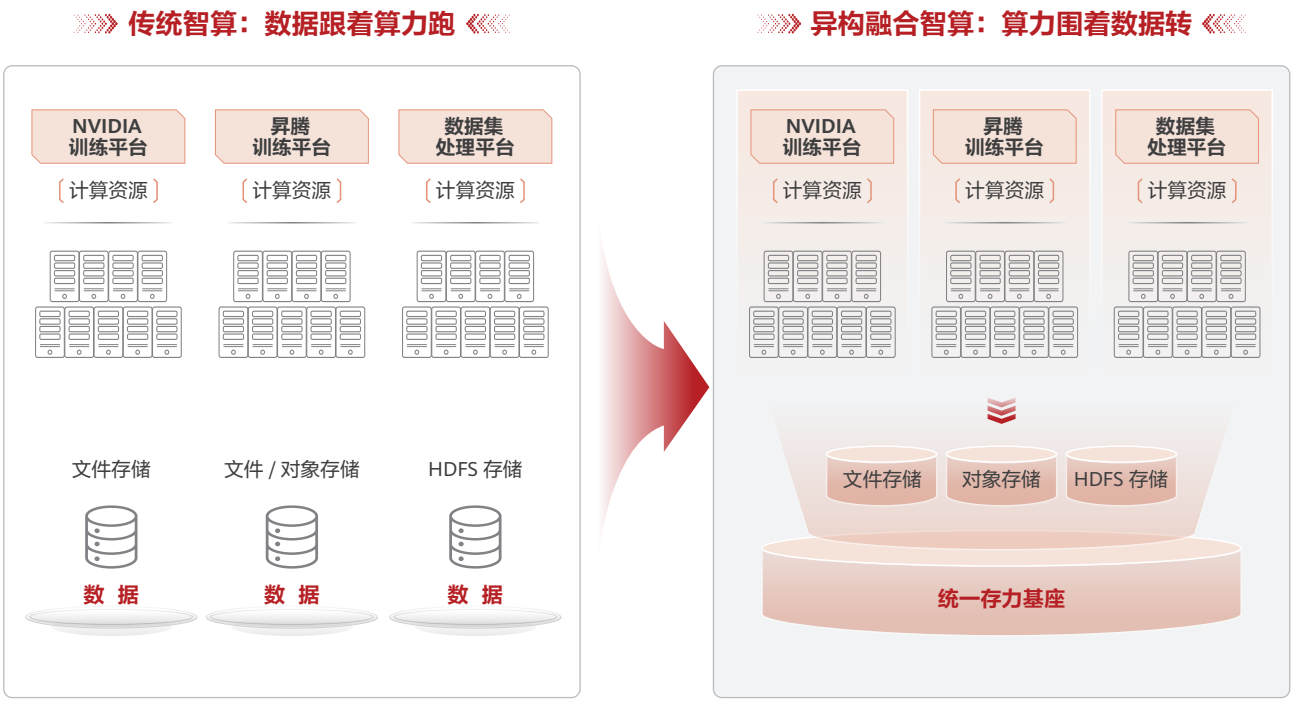


图 3 - 算力中心架构到存力中心架构

1. 先进介质：全面走向闪存，加速数据存取速度

传统的机械硬盘已经无法满足快速访问和处理大规模数据的需求，而固态硬盘在读写性能上远超机械硬盘。单个固态硬盘的 IOPS 提升千倍，同时还具有低延迟和大吞吐量优势，能够更好地适应大模型对数据读写的高吞吐、低时延需求。数据读写性能的大幅提升将减少计算、网络等资源等待时间，加速大模型的研发与应用。此外，随着存储颗粒类型和堆叠层数的突破，固态硬盘单价持续下降，使得全闪存存储建设成本变低，成为 AI 大模型的理想选择。

2. 先进架构：统一数据底座，承载 AI 全流程业务

AI 业务流主要包括数据获取、数据预处理、模型训练和推理应用，为了实现高效的数据共享和处理，需要采用多种不同的非结构化协议，如 NFS 协议、S3 协议、HDFS 协议和 POSIX 高速访问接口等。传统解决方案是采用多种存储协议，由于需要在不同系统间来回拷贝数据，会严重影响数据处理效率，浪费存储空间，增加运维难度。因此，建设数据易共享、高性能、易扩展的统一数据底座来承载 AI 全流程业务，是最好的选择。这样可以提高数据处理效率，减少存储空间浪费，降低运维难度，同时支持多种非结构化协议的使用，满足不同场景下的需求。

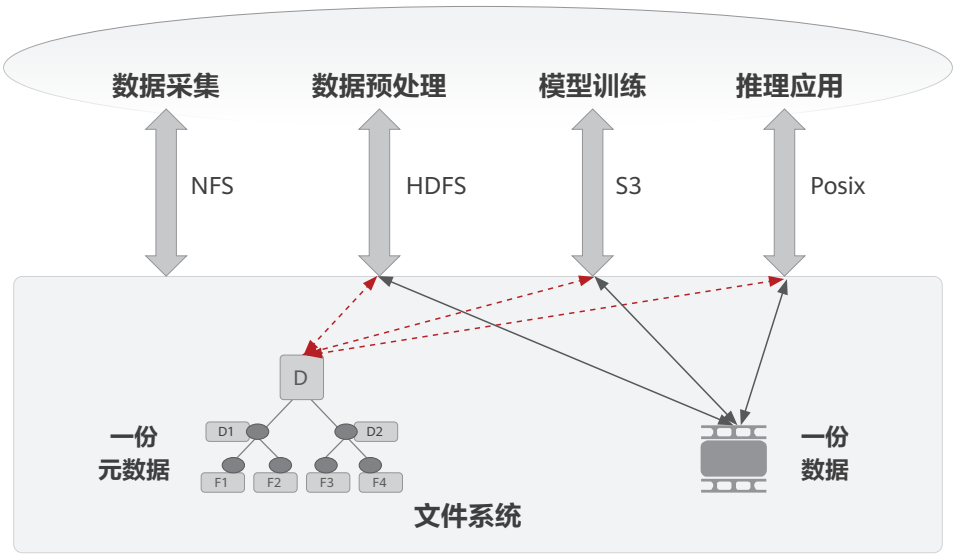


图 4 - 多业务接口的统一存储逻辑图

先进架构应具备无损多协议互通和数据全生命周期管理的能力。

无损多协议互通，数据访问透明

统一存储可以承载 AI 全流程业务，并且兼容 AI 全流程工具链所需的 NAS、大数据、对象和并行客户端等协议。同时，该系统要保证各协议的语义无损，达到与原生协议一样的生态兼容性要求。此外，该系统还需要具备高效的数据流转能力，以便在不同阶段实现数据零拷贝和格式零转换，从而确保前一阶段的输出可以作为后一阶段的输入，并实现 AI 各阶段协同业务的无缝对接，达到零等待的效果。

统一命名空间，数据全生命周期管理

首先，AI 存储系统需要同时具备高性能层和大容量层，以便满足不同应用场景的需求。其次，它应该对外呈现统一的命名空间，方便用户管理和访问数据。此外，该系统还需要具备数据全生命周期管理的能力，包括指定数据首次写入时的放置策略、设置丰富的数据分级流动策略以及数据预取能力等。其中，数据分级流动策略可以根据数据的访问频度和时间进行设置，以提高系统的性能表现；数据预取能力可以通过预热策略来加速计划性任务的冷启动速度，提高系统的响应效率。

3. 先进设计：并行文件系统，提升数据访问效率

多模态大模型对 AI 数据存储系统提出了多样性要求，如数据获取、预处理、训练和推理的过程中，需要存储系统能提供并行文件系统以提高 AI 芯片的工作效率。并行文件系统应具备以下能力。

- 一是并行文件系统需要支持高并发、高带宽、高 IOPS，以提高 GPU/NPU 的训练推理效率。同时并行文件系统需要具备高扩展性，能够支持 EB 级的数据量，且性能随节点数增加而线性增加，保证数据可以均衡分布在所有存储节点上，业务压力均衡到各节点，实现访问无瓶颈，保障规模扩展场景下的系统性能。
- 二是并行文件系统需要近计算加速。存储为计算节点提供高性能的并行客户端，让数据更靠近计算。在并行客户端上可以构筑丰富的功能，来加速 I/O 性能。首先可以通过智能预取算法将数据提前预取到计算节点本地的高速缓存，同时还可以将各种数据聚合后再传输到存储；其次客户端不用通过协议服务器，可以直接访问所有存储节点传输数据，从而避开 I/O 路由瓶颈和性能损耗；另外客户端支持多链路数据均衡方式，可以极大提升计算节点获取数据的能力。
- 三是分布式并行训练需要网络具备无丢包、低时延、高吞吐的能力。采用远程直接内存访问技术（RDMA）进行 I/O 数据交换，相比传统 NFS 使用的 TCP 协议，减少了 CPU 负担和协议开销，可实现高带宽、低时延和低资源消耗的效果，RDMA 已成为人工智能计算中心主流选择。

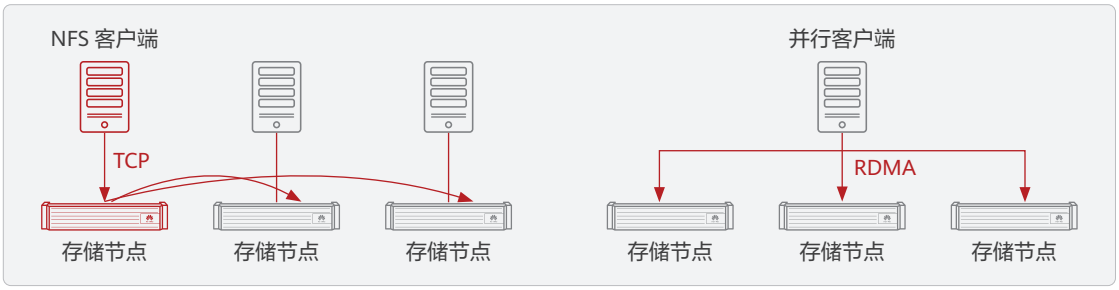


图 5 - 传统文件系统架构向并行文件系统架构发展

- 四是大模型在训练完成后到实际应用中存在一定的滞后性，为了解决这个问题并提高模型应用的及时性和准确性，可以采用将不同格式的文件数据（如 Word、Excel、PDF、图片等）切片后更新到向量数据库的方式。这样，在用户提问时，可以通过检索与问题相似度高的向量，并将提炼后的 Prompt 输入到既有大模型中进行推理，从而实现通用大模型在行业的快速落地。已生成的推理结果会同时记录在缓存中，供下次提问时快速返回。这种结合向量数据库的方法有助于将行业最新知识应用于既有大模型推理，提高大模型的准确性和实用性。

4. 安全可靠：存储内生安全，保护企业核心资产

存储内生安全是一种新兴的数据安全理念，它强调在数据存储系统中内置安全机制，以保障数据的机密性、完整性和可用性。相比于传统的外部安全措施，存储内生安全具有更高的灵活性和可控性，能够更好地适应不断变化的安全威胁。

在大模型的应用中，存储内生安全通过将安全功能融入到数据存储系统中实现对数据的保护。这种方式可以避免将敏感信息暴露在外部网络中，从而降低数据被攻击的风险。同时，存储内生安全还能够提供灵活的安全管理策略，根据不同用户的需求进行定制化的安全设置，提高数据的安全性和合规性。

作为数据的最终载体，存储设备必须具备安全可靠的内生能力，以增强整个大模型系统的数据防护能力，构建数据安全最后一道防线。从底层硬件角度，存储内生安全主要包括构建关键硬件自主能力、硬件三防（防侧信道、防故障注入、防物理攻击）和可信启动等技术。从软件算法角度，重点解决开源软件的风险治理，数据存储应采用 AIR GAP 技术来保障数据安全传输、一写多读（WORM，Write Once Read Many）技术来防止文件被篡改、病毒侦测分析来预防被病毒勒索、执行环境提前检测来确保数据可信、数据访问的全路径和内存加密技术防止数据被泄露。从应用效果角度，通常采用存储加密及定期备份两种方案面对突发情况下数据丢失或破坏等问题，使数据的安全性有充足的保障。

5. 绿色低碳：存储高效节能，建设绿色存力底座

在全球范围内，节能减排已成为共同使命，各行各业都在积极追求“碳中和、碳达峰”的目标。为了建设可持续发展的运营商人工智能计算中心，构建高效节能的存储设施至关重要。

碳足迹（CFP，Carbon Footprint of Products）是指产品和服务整个生命周期过程引起的温室气体排放的集合。对于存储产品来说，降低能耗需要从碳足迹入手，深入到产品全生命周期的各个环节。为了推动存储产品的节能减排，需要采取有针对性的碳减排设计和绿色节能技术。

AI 先进存力的绿色低碳化需要全面的存储能力提升，主要体现在以下几个方面。

（1）高密存储硬件，提高单位物理空间的存储密度

普通通用型服务器受限于散热和空间限制，硬盘数量有限，通常为 1U10 盘、2U24 盘、4U36 盘等。通过设计高密存储型节点，可以做到 1U32 盘、2U36 盘、5U80 盘、4U80 盘、5U120 盘，密度达到传统存储的 2 到 2.6 倍。这种高密存储型节点相对于通用型服务器，减少了节点 CPU、内存及配套交换机的使用，同等容量下能够节约 10% 到 30% 的能耗。

（2）大比例 EC，提高存储空间有效利用率

在 AI 场景中，存储资源利用数据纠删码 (EC，Erasure Code) 技术替代了传统的三副本方案，使用 N 份数据加 M 份校验的组合可以达到甚至超过 3 副本的可靠性。采用大比例的 EC 机制可以有效提升存储空间的利用率，例如当 EC 为 2+2 时，存储空间的利用率可以由 33% 提升至 91%。

（3）数据缩减技术，节约存储空间

通过数据缩减技术，可以在不失真情况下将实际需要保存的数据量大幅减少，以更少的物理容量存放更多的数据，从而降低存储设备的能耗。目前业界能够在数据库、桌面云、虚拟机等业务场景实现 2 到 3.6 倍的数据缩减率，从而节省 50% 以上的能耗。这种技术可以有效地提高存储设备的效率和可靠性，同时也减少了企业的运营成本。

（4）硅进磁退，提高全闪存占比，降低能耗

在相同容量下，固态硬盘相比机械硬盘能够降低 70% 的能耗，同时节省 50% 的空间占用，同时，固态硬盘具有高密度、高可靠、低延迟和低能耗等特点。通过大规模部署全闪存，可以大大降低人工智能计算中心的能耗，帮助运营商实现绿色节能和可持续发展。

（5）存储全生命周期管理，推进绿色节能

绿色节能需要贯彻存储产品的全生命周期，包括原材料选取、制造、运输、使用到最终废弃。在生产制造环节，制造工厂使用光伏发电，选择铝、锡等可再生材料，同时采用“零波峰焊接”技术和标签无纸化，以降低能源消耗和减少环境污染。在存储产品生命周期末端，建立完善的回收系统，以环保的方式处理电子废物，实现最大化的循环再利用，以减少对环境的影响。这样的绿色节能措施可以帮助存储产品实现更可持续的生产和消费方式。

5 运营商 AI 先进存力发展建议

为充分使能运营商提供万物互联的链接优势，服务千行百业的行业优势，催生新产业新模式，持续做强做大我国数字经济，本白皮书针对我国运营商 AI 先进存力发展，提出如下建议：

» **一是重生态**，未来 AI 的竞争是全栈能力的竞争，是生态之争，而运营商作为整个 AI 生态链中的核心节点，具备加强全栈能力，做大生态的天然优势。AI 生态的构建涉及存储层、网络层、计算层、资源层、框架层、大模型和智能应用层等，这些决定了 AI 未来发展的根本。大模型是AI时代的“腰杆”，而生态是大模型发展的关键。大模型研发需要各个层级生态的紧密配合，缺一不可。当务之急是构建一个大模型生态，这才是未来大模型运营服务和智能应用的强大底座。

运营商要打造先进算力、存力、运力布局，以基础大模型为桥梁，构筑基础大模型的多模态能力，带动多模态算法研发及云边端芯片研发，形成自主可控的人工智能平台。在此基础上，以先进人工智能计算平台为底座，开发多行业多场景大模型，对内服务公司智慧化运营，对外赋能行业数智化转型，并最终形成汇融算力、存力、运力及数据资源的国家级人工智能计算体系。同时，打造人工智能基础模型开源技术体系，助力产出规模和性能国际领先的大模型，支撑工农业生产、社会治理、人民健康等应用创新。

- **二是重发展**，在运营商 AI 存储领域加快先进存力应用，将数据存储打造为继 5G 后的国家信息技术新名片。倪光南院士指出，在多年自主创新的基础上，我国数据存储产业取得了重大发展，基本上达到了科技自立自强的要求，例如构筑在我国自主核心数据存储芯片、系统架构和软件之上的存储整机和系统已具备国际领先技术水平，其中华为存储在 2022 年全球存储市场排名第二，并进入国外金融行业等关键市场，成为国际头部存储整机厂商，联想、新华三和浪潮在 2022 年的全球市场占比也进入了前 10 名。目前国产品牌厂商在国内市场占比 85%，在全球市场占比 20%，在海外有巨大增长空间。

运营商作为 AI 产业的主力军，有强大的能力来牵引先进存力的发展。能力越强，责任越大，建议运营商在新建 AI 基础设施时加大安全可信的先进 AI 存力应用，推动我国存储产业跨越式发展。

- **三是重安全**，充分重视 AI 存储的自主可控和安全可信能力，并设定相应标准，严格执行。当前我国存储产业还存在较大的安全隐患，首先硬件不自主，当前运营商采购中机械硬盘约占 80%，全部来自美日两国的 3 家供应商，存在非常大的安全和供应风险；其次软件不可信，国内 95% 以上分布式存储厂商采用国外开源软件，如 Ceph、Luster、Gluster、DAOS 等，多数简单修改后商用，没有掌握核心技术。在运营商这样重要关键信息基础设施中大规模使用国外软硬件，潜在安全风险非常高。

为了加强运营商 AI 存力的安全性，建议率先在运营商实现公平公正的安全测评，重视对数据存储产品的安全审查，包括供应链安全、信息安全、数据安全等。尤其是对于数据存储整机、主控芯片、存储文件系统等关键核心技术，建议根据第三方测评机构给出的相关结论作为选择依据。

参考文献

1. Bishop, C. M. (2006).《模式识别与机器学习》，Springer. ISBN 978-0-387-31073-2.
2. 我国人工智能大模型发展动态 http://www.miitnet.com/news/7005.html
3. 《AI 智算基础设施架构研究及关键技术分析》
4. 中国移动发布“九天”大模型，https://www.sohu.com/a/696201560_121010226
5. 美中科技战从技术管控到投资管控 高科技背后的博弈与风险，https://www.bbc.com/zhongwen/simp/business-66530604
6. 柯瑞文：中国电信将发布星河通用视觉大模型 2.0，https://www.21jingji.com/article/20230428/herald/a9da1eec3ef897c91f50fa493b642209.html
7. 鸿湖图文大模型 1.0：中国联通发布面向运营商增值业务的图文生成模型，https://www.sohu.com/a/692475378_121745009
8. 中国移动 NICC 新型智算中心技术体系白皮书

汇 聚 产 业 新 动 能 · 共 创 A I 新 时 代