中国移动
China Mobile

中国移动NICC新型智算中心技术体系白皮书

CHINA MOBILE NEW INTELLIGENT COMPUTING CENTER TECHNOLOGY FRAMEWORK WHITE PAPER

中国移动通信集团有限公司
2023年8月

前 言

ChatGPT 系列大模型的发布，不仅引爆全球科技圈，更加夯实了人工智能（Artificial Intelligence, AI）在未来改变人类生产生活方式、引发社会文明和竞争力代际跃迁的战略性地位。

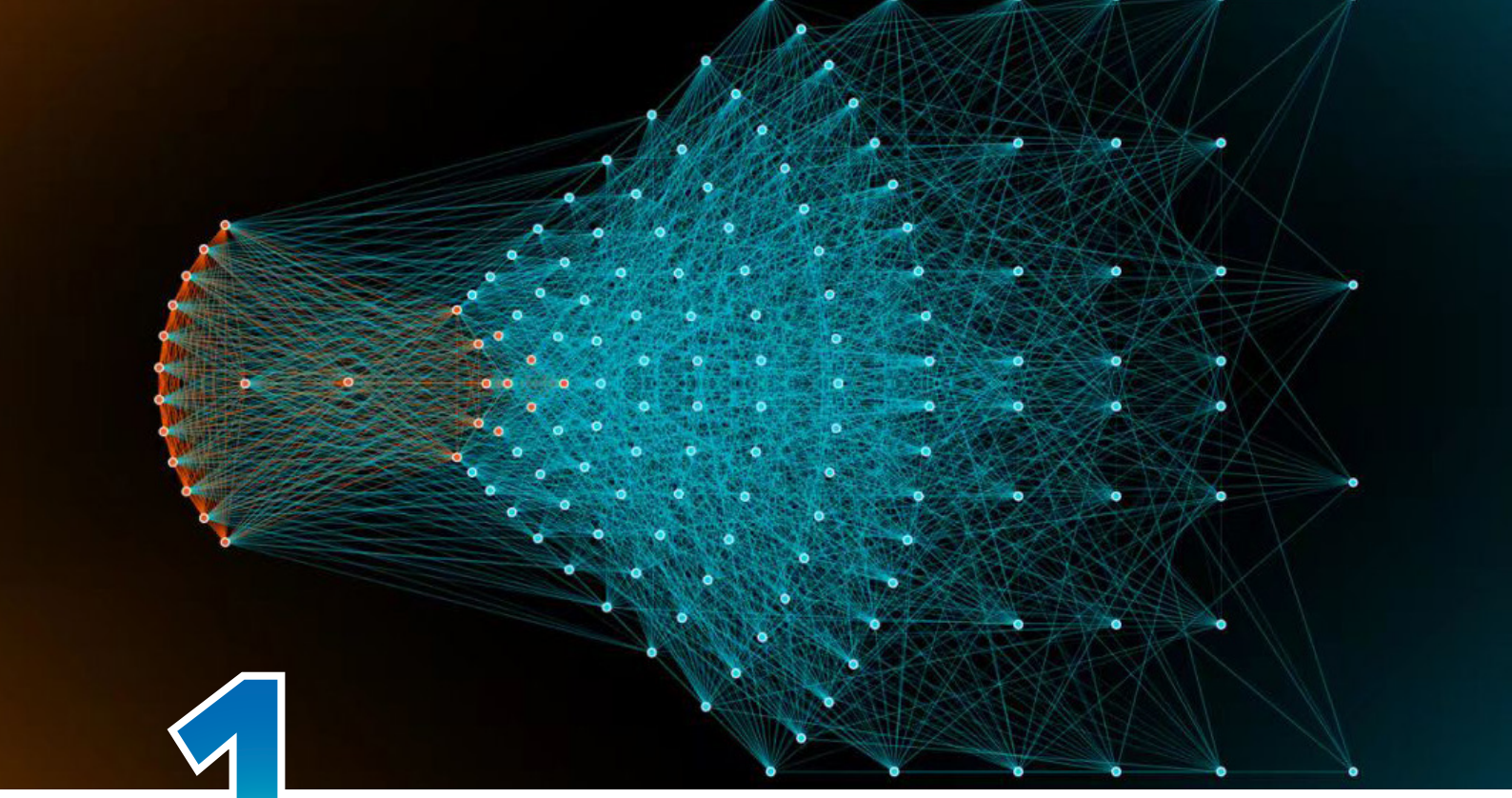
当前各国政府已全面布局 AI，作为 AI 技术发展的关键底座，智算中心的建设和部署在全球范围内提速。然而，早期建设的智算中心，以承载中小模型、赋能企业数智化转型为主要目的，在技术标准、生态构建、业务发展和全局运营等各方面仍有待提升。当追逐大模型成为行业标准动作，面向大模型的新型智算中心（New Intelligent Computing Center，NICC）成为新时期关注的焦点。

新型智算中心的建设是一个系统工程，需要“算存网管效”多个维度的协同设计。中国移动从自身战略转型出发，为构建智能服务的核心和基础，定义新型智算中心技术体系架构，并面向未来大模型孵化，从新互联、新算效、新存储、新平台和新节能等五个领域提出下一代技术演进建议，希望本白皮书能够为合作伙伴在新型智算中心的硬件设备选型、算力集群设计、机房散热规划、软硬工程调优、全局运营调度等多个方面的技术路线选型提供帮助。

本白皮书在中国移动集团有限公司技术部和计划建设部指导下，由研究院牵头编写，期间得到了来自华为、浪潮信息、新华三、曙光、超聚变、中兴、寒武纪、燧原、壁仞、趋动科技、星网锐捷、昆仑芯、天数智芯、盛科、云合智网、云豹智能、云脉芯联、星云智联等多家企业的大力支持。

新型智算中心技术体系的构建与成熟需要产业链各方凝聚共识，明确行业应用和服务的共性要求，中国移动希望同行业一道，共同推动智算关键技术成熟，共同繁荣国内 AI 生态发展。

目录 contents



1 智算中心行业发展现状

1.1 智能算力跃升为全球第一大算力，智算中心建设如火如荼

1956 年第一次 AI 发展浪潮信息伊始，60 多年来，从理论探索到大数据驱动，从深度学习到大模型智能涌现，AI 正在成为一项新兴的通用型技术，向多场景、规模化、AIGC（AI Generated Content）等方向快速演进。智能算力作为 AI 的底座型技术迎来需求井喷。据统计，到 2030 年，全球智能算力需求增长约 390 倍，增速远超摩尔定律。据《中国算力发展指数白皮书（2022）》指出，我国智能算力也在近几年保持快速增长态势。2021 年我国智算规模已达到 104E FLOPS，占比超过总算力的 50%，预计到 2030 年将升至 70%，成为算力的主要增长极。智算成为全球第一大算力已是大势所趋。

随着 AI 在赋能产业发展、促进数实融合方面发挥出愈加显著的作用，各国政府纷纷发布政策引导其发展。美国为加强其在 AI 领域研发和部署的领导地位，于 2019 年签署《美国人工

图 1-1 美国智能超算中心

智能倡议》（American AI Initiative）[1]，旨在从国家战略层面重新分配资源，用于 AI 研发，以应对来自“战略竞争者和外国对手”的挑战。之后，在 2021 年颁布《美国创新与竞争法案》[2]，高度关注 AI 与机器学习、高性能计算、半导体等十大关键技术领域。欧盟在 2021 年发布《2030 数字指南针：欧洲数字十年之路》[3]，要求到 2030 年 75% 的欧盟企业使用云计算、大数据和 AI 技术。我国也高度重视 AI 技术发展，自 2017 年以来国家各部委和地方政府相继出台政策，指导 AI 产业发展规划，鼓励企业加大人才引进和研发力度，并明确指出要积极推动智算中心有序发展。**至此，智算中心作为一种新型算力基础设施为大家所熟悉。不同于传统的云数据中心和超算中心，智算中心是以 GPU、AI 加速卡等智能算力为核心、集约化建设的新型数据中心，为人工智能应用提供所需的算力服务、数据服务和算法服务，使能各行各业数智化转型升级。**

智算中心的战略地位不断提升，为构造未来竞争发展优势，很多国家都在积极开发和部署智算中心。其中，美国能源部及国家科学基金会主导，将智算中心和超算中心结合，建设超大规模智能超算中心，为科学研究提供高性能计算资源（如图 1-1），例如，橡树岭国家实验室的 Summit（3.4E）[4]，阿贡国家实验室的 Polaris 和 Aurora（约 10E）[5]，劳伦斯伯克利实验室的 Perlmutter（3.8E）等，这些智能超算中心往往具有单体算力大、技术领先等特点。美国科技巨头也是智算中心的主要建设者，包括谷歌的开放机器学习中心（9E），特斯拉 Dojo 集群（据称 2024 年末规模达到 100E），Meta AI 超级计算机（9.9E）等。

图 1-2 国内部分智算中心

国内智算中心建设热潮始于 2020 年，目前已有 40+ 城市建成或正在建设智算中心（如图 1-2），包括武汉人工智算计算中心（200P）、南京智能计算中心（800P）、合肥先进计算中心（12P）、鹏城云脑 II（1E）等，其中 12 个位于“东数西算”八大枢纽，这些智算中心主要由地方政府与企业合建，总体投资规模超千亿，旨在带动当地产业智能化升级。国内互联网和 AI 企业自建的智算中心是国内智能算力的重要组成，如阿里在张北和乌兰察布建设的总规模达 15E 的智算中心，旨在结合智能驾驶、智慧城市等业务，探索云服务后的智算服务新业态；百度在山西阳泉建设规模 4E 的智算中心，孵化国内首个正式发布的大模型“文心一言”；商汤作为国内头部 AI 企业，投资 56 亿在上海临港建设人工智能计算中心，规模超 4E，主要面向智慧商业、智慧城市、智慧生活和智能汽车四大板块，发展 AIaaS（AI as a Service）服务。

1.2 早期智算中心在技术、标准、生态、运营等方面仍面临挑战

当前智算中心主要以单供应方全栈体系构建为主，尚未形成业界统一的设计方案，因此各地智算中心在技术、标准、生态、运营等方面仍面临挑战。

在技术方面

早期建设的智算中心以承载中小模型为主，AI 服务器大多是 PCIe 机型，配备独立的文件存储，互联方式则以节点内 PCIe 通信与节点间传统以太网为主。随着通用大模型的普及，智算中心的设计思路需要从原先以单芯片、单服务器粒度提供算力服务的模式，转变为支持巨量并行计算，提供高吞吐、高能效的集群算力。

在标准方面

由于各地智算中心大都是当地政府与 AI 芯片、整机厂家合作建设为主，技术方案深度绑定，容易形成多种派系。亟需通过制定行业标准，一方面降低客户学习和使用的时间成本，另一方面加强产业链上下游企业的协同，促进智算产业的高质量发展。

在生态方面

因为 AI 是软硬深度耦合的技术栈，国外主流产品“先入为主”，主导生态发展，相比之下国内 AI 起步较晚，在芯片算力和软件栈适配方面均存在差距。在智算生态竖井式发展的当下，需要加强引导，为后续 AI 应用的适配和跨架构迁移奠定基础。

在运营方面

各地智算中心的服务对象多为区域内的行业客户、科研院所和高校，较少考虑全局协同，随着东数西算、东数西渲等应用需求不断丰富，需要提前布局跨区域的全局算力调度，提升算力高质量供给和数据高效率流通。

由此可见，未来智算中心亟需朝着技术先进、标准统一、软硬协同、兼容开放的方向发展。

2

NICC 新型智算中心技术体系架构和发展路径

2.1 NICC 新型智算中心技术体系架构

结合大模型技术的发展趋势以及对智算中心建设和使用现状的分析，我们认为 ChatGPT 等预训练大模型的出现，必将带来 AI 基础设施的变革，传统的算力堆叠方式已然失效，智算中心需要在**互联、算效、存储、平台、节能**五大领域进行系统化的重构，才能支撑起大模型对千行百业的革新与改造。为此，中国移动结合自身转型战略和一线客户需求，提出 NICC 新型智算中心（New Intelligent Computing Center）。

区别于早期建设的智算中心，NICC 新型智算中心是以**高性能** GPU、AI **加速卡**等集群算力为核心，集约化建设的 E 级超大规模算力基础设施，具备从硬件设施到软件服务的端到端 AI 全栈环境，支撑超大规模、超高复杂度的模型训练和推理业务，最终赋能行业数智化转型升级。

NICC 技术体系由“三层两域”构成（如图 2-1），分别是基础设施层、智算平台层、应用使能层、智算运维域和智算运营域。其中基础设施层提供计算、存储、网络等硬件资源；智算平台层作为资源管理的核心，提供裸金属、虚机和容器等多样化实例以及细粒度的资源池化能力，在此之上搭建算力原生平台提供应用跨架构迁移能力；应用使能层集成行业主流 AI 开发框架以供应用开发调用。智算运维域主要负责对底层 IaaS（Infrastructure as a Service）资源进行管理维护，确保系统的稳定运行；智算运营域对接外部客户，提供计量计费、访问、交易等界面，对内根据上层任务进行资源编排调度。

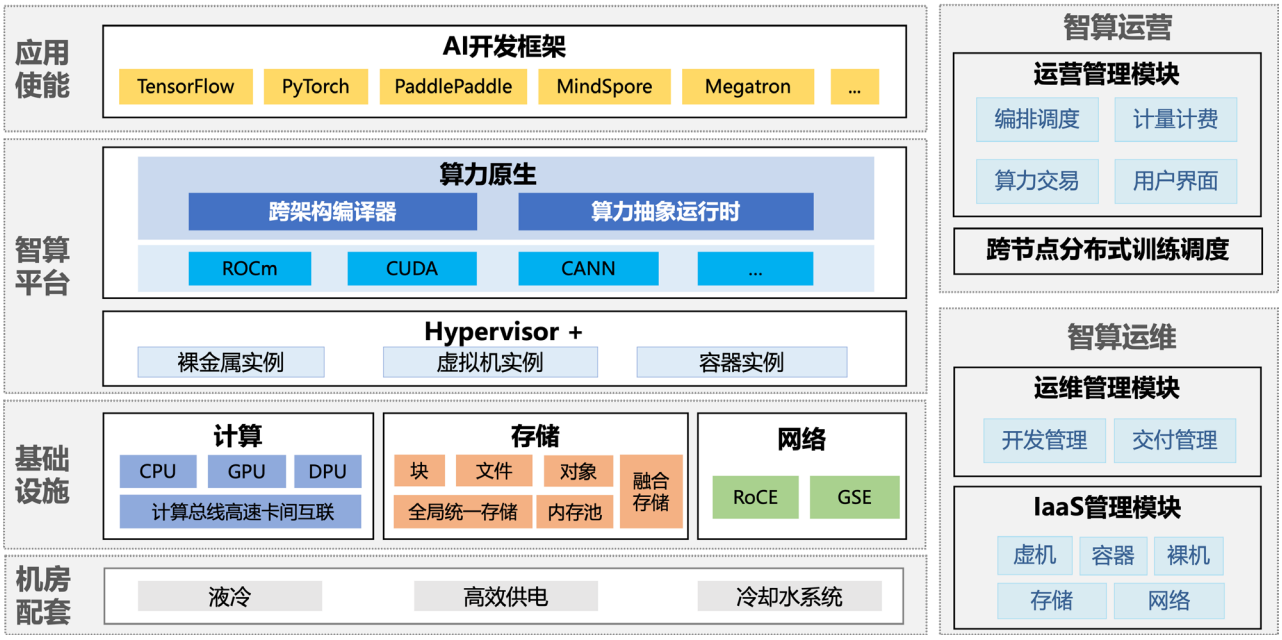


图 2-1 新型智算中心技术体系架构

2.2 NICC 新型智算中心技术发展路径

为释放智能算力极致性能，NICC 的设计方案既要考虑计算、存储、网络三大维度的横向协同，也要兼顾软件平台与硬件资源的纵向协同，同时锚定技术先进、标准统一、软硬协同、兼容开放的目标，广泛且高效地支撑智能化应用场景。我们认为 NICC 的发展将分为两个时期：

1) 集群时期： 这个时期最显著的特征是数据及模型出现巨量化趋势，千亿级的模型已经出现，对智算底座的算力能力和扩展性均提出高要求。在设备形态方面，GPU、AI 芯片以扣卡模

组为主，服务器形态多为单机 8 卡，DPU 按需引入解决裸金属管理、存储加速等业务痛点；硬件资源开始按照集群的方式部署，相比提升单芯片算力，芯片间的高速互联方案落地更为关键。互联方案以服务器节点为界限，节点内外高速互联技术各自发展，节点内采用高速计算总线，节点间采用 100G/200G 高速无损网络；在存储方面，原先独立部署的文件、对象存储逐渐向融合存储演进，提升数据交互效率；平台应具备池化算力分配能力，实现底层智算资源的细粒度分配、动态调度和一体管理。分布式并行训练框架需要引入提升模型训练效率。为配合高算力需求，散热系统逐渐从风冷向冷板式液冷过渡。

2) 超级池化时期：当大模型迈进万亿参数量规模，算力、显存和互联的需求再次升级，智算中心将真正进入超级池化时代，高速互联的百卡组成的“超级服务器（Super Server，S^2）”可能成为新的设备形态。传统以单机 8 卡为最小单元的智算中心设计思路需要革新，“超级服务器”内需要打造统一的协议实现 CPU、GPU、AI 芯片、显存、存储等池化资源的无缝连接，进而通过 GSE 等高性能交换网络，达到极高吞吐、极低时延的系统算力；为推动算效能力进一步提升，基于存算一体架构的大算力芯片将开始逐步应用；存储系统在“超级服务器”内支持内存池技术，对外扩展支持全局统一存储；针对日益割裂的智算生态，需要构建基于算力原生平台的跨架构开发、编译、优化环境，屏蔽底层硬件差异，从软件层面最大化使能异构算力融通。散热系统方面，为匹配“超级服务器”设施发挥出最大算力能力，浸没式液冷也将逐渐规模落地。

我们认为，新型智算中心当前已处在“集群时期”，中国移动和部分企业已经按照集群的思想构建 AI 基础设施；面向中远期，我们应重点攻关“超级池化时期”的关键技术，尽快形成行业共识，加速相关核心技术和产业成熟。

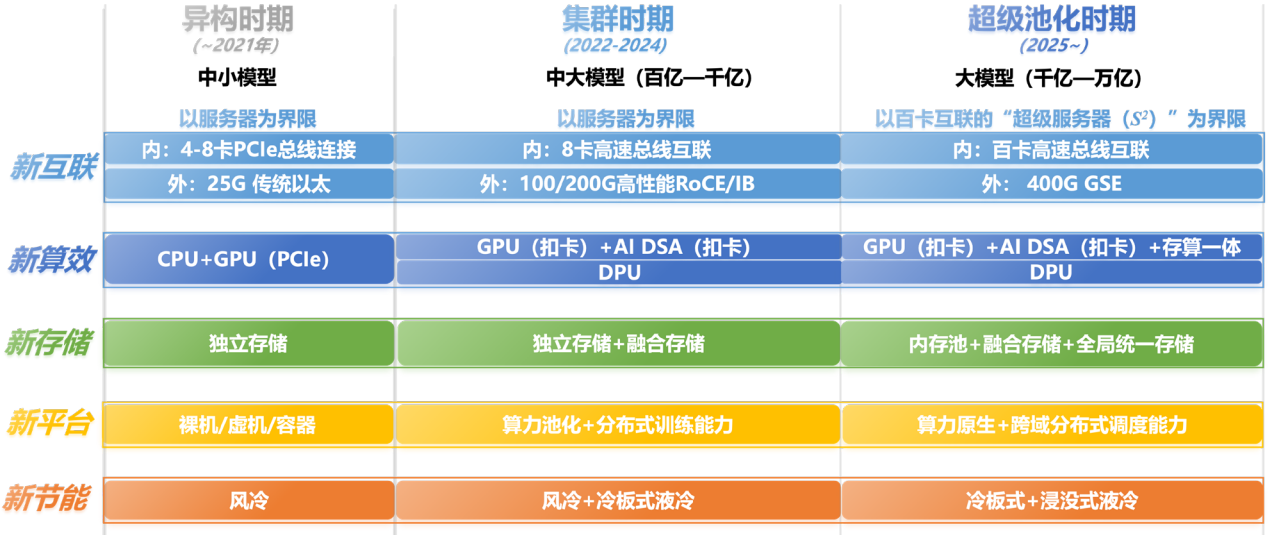


图 2-2 新型智算中心技术发展路径

3 新互联——打破算力瓶颈

3.1 集群内的高速卡间互联

大模型浪潮除了带来算法及软件革命，也拉开了 AI 基础设施变革的序幕。一方面，算法结构的创新影响了 AI 芯片在算力精度范围和专用加速电路等方面的设计，但单芯片算力提升的速度仍无法赶上模型参数的发展速率（如图 3-1）；另一方面，由于巨量参数和庞大的数据样本，模型的尺寸已经远超出单个 AI 芯片甚至单台服务器的计算能力，亿级以上的模型需要部署在高速互联的多个 AI 芯片上，分布式并行训练。当前，相较于单芯片能力提升，多芯片集群的规模化能力及效率是产业研究的重点。

3.1.1 大模型分布式训练需要高速卡间互联

在大模型迸发出知识涌现能力之前，AI 的主流场景是中小模型承载的计算机视觉类（Computer Vision，CV）应用，模型参数在亿级以下，如 ResNet50（~25M）等。此类模

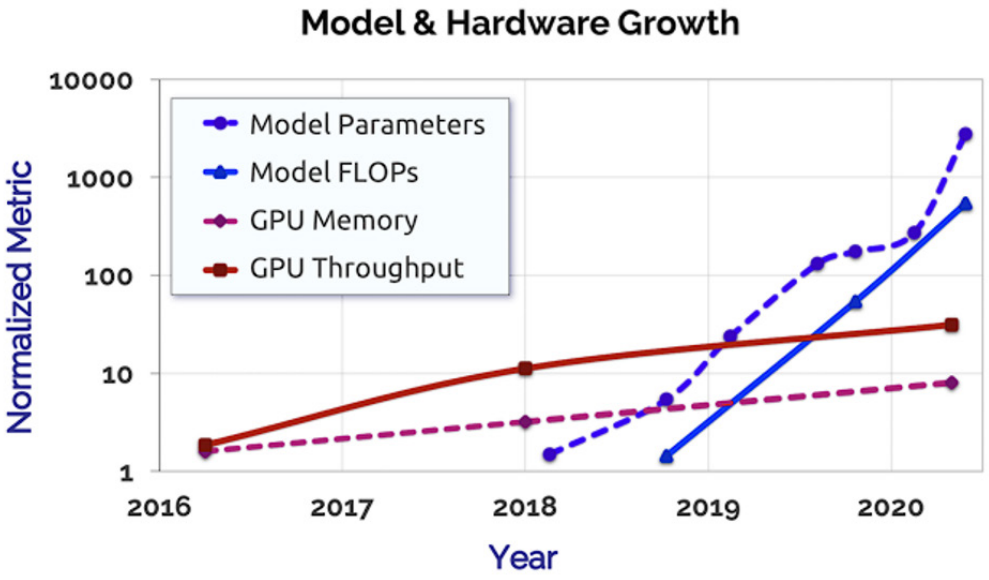


图 3-1 模型参数量和 GPU 算力的发展趋势 [6]

型显存占用集中在单卡或者单服务器节点，训练模式多采用单卡运行或单节点内多卡数据并行，每张卡或节点上都有完整的模型结构，卡间通信主要传输训练数据，因此通信需求不频繁，带宽要求在几十 GB 级别，传统的 PCIe 设备形态即可满足其训练需求（如图 3-2）。

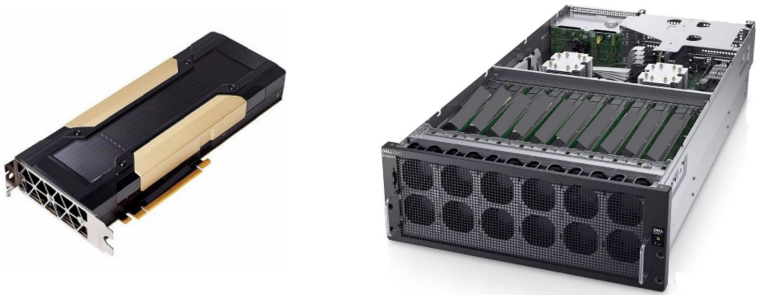


图 3-2 PCIe 形态的插卡和整机设备

当模型参数量迈进千亿规模，如 GPT3（1750 亿），训练模式也从单芯片运行转变成多芯片分布式运行，数据样本和模型结构被切分到多张卡或者节点上，卡间或者节点间不仅有数据样本的通信，还有模型梯度的频繁传递，对卡间的互联能力在带宽和拓扑结构两大方面产生高要求。

常用的分布式并行策略主要分为数据并行（Data Parallel，DP）和模型并行（Model

Parallel，MP）两大范畴，两者通信操作不同，对卡间的带宽和互联拓扑要求也不同：

● 数据并行的实现思路是每个计算设备上（每张卡或者节点）都有一个完整模型，将数据集拆分到多个计算设备上同时训练，在反向传播中，各个设备上的梯度进行归约操作求平均（AllReduce），再更新模型参数。通信操作中主要使用到 Ring-Allreduce 算法，多个计算设备采用环状互联拓扑，通信带宽要求多为几 - 几十 GB/s。

● 模型并行主要分为流水线（Pipeline Parallel，PP）和张量并行（Tensor Parallel，TP），其中流水线并行最早由谷歌在 Gpipe 算法 [7] 中提出，将模型按照层的维度拆分成多个 Stages 放在每个计算设备上，训练过程是逐层顺序计算，通信数据量比数据并行小，对拓扑无特殊要求，点对点互联即可，通信带宽要求在几 - 十几 GB/s；张量并行由英伟达在 Megatron-LM 论文 [8] 中提出，将模型在层内进行切分，训练过程中前向和反向传播中都涉及 Allreduce 操作，通信量大且频繁，计算设备通常要求是全互联（Fully connected，FC）甚至交换拓扑（Switch），带宽需求在几百 GB/s。

表 3-1 不同的分布式并行策略及对应的卡间互联要求

策略	通信模式	互联拓扑，带宽需求
数据并行 DP	Allreduce	环状或全互联，常规需求，几 ~ 几十 GB/s
流水线并行 PP	P2P	点对点相连，常规需求，几 ~ 十几 GB/s
张量并行 TP	Allreduce	环状或全互联，带宽需求高，几百 GB/s

由于大模型训练对芯片互联提出高带宽、低延时以及拓扑结构高扩展性等要求（如表 3-1），PCIe 形态设备在通信带宽和模式上都难以为继。在带宽方面，PCIe 4.0*16 最高为 64 GB/s，无法覆盖百 G 带宽需求；在通信模式方面，卡间通信必须经过 CPU 绕转甚至跨 CPU NUMA，不仅带来通信延迟，还增加 AI 算法开发难度；在扩展性方面，部分厂家曾采用桥接器搭配自研的通信协议实现卡间高速互联，但因整机主板设计和桥接器的机械应力限制，互联数量基本在 4 卡及以下，扩展能力有限。因此，PCIe 设备形态逐渐被扣卡模组形态的产品（如图 3-3）替代，成为业界大模型训练的主流解决方案。

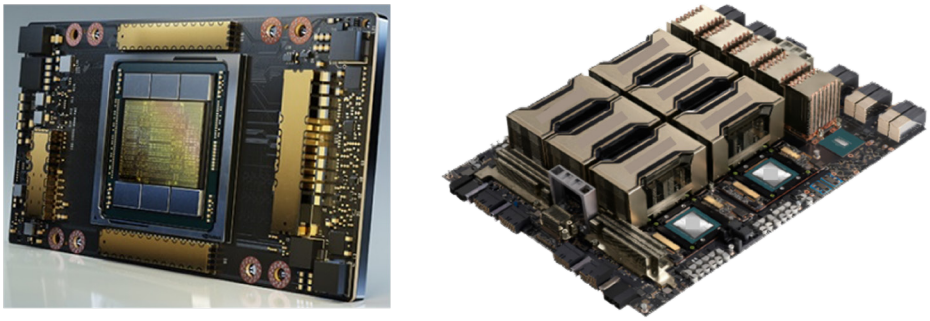


图 3-3 GPU 扣卡模组形态

3.1.2 “七国八制”的卡间高速互联技术现状

针对中小模型训练，基于 PCIe 设备形态的解决方案已经非常成熟，面向大模型场景，基于扣卡模组的卡间高速互联方案则呈现“七国八制”的局面。目前行业主要分为私有和开放技术两大类。私有方案以英伟达 NVLink 为代表，目前已经发展到第四代（如图 3-4）。第一代到第二代的演进主要体现在互联拓扑的转变，从 cube 直连演变为 Switch 交换拓扑，第三代在交换拓扑的基础上，通过增加单卡的 NVLink 通道数提升点对点（Peer to Peer, P2P）带宽，第四代则通过完善多种协议内容，进一步实现 C2C（chip to chip）、AI 卡间以及服务器节点间的统一连接，达到最高至 900GB/s 的 P2P 带宽，以及 256 个 H100 的全互联能力，极大地提升了大模型并行训练的效率。

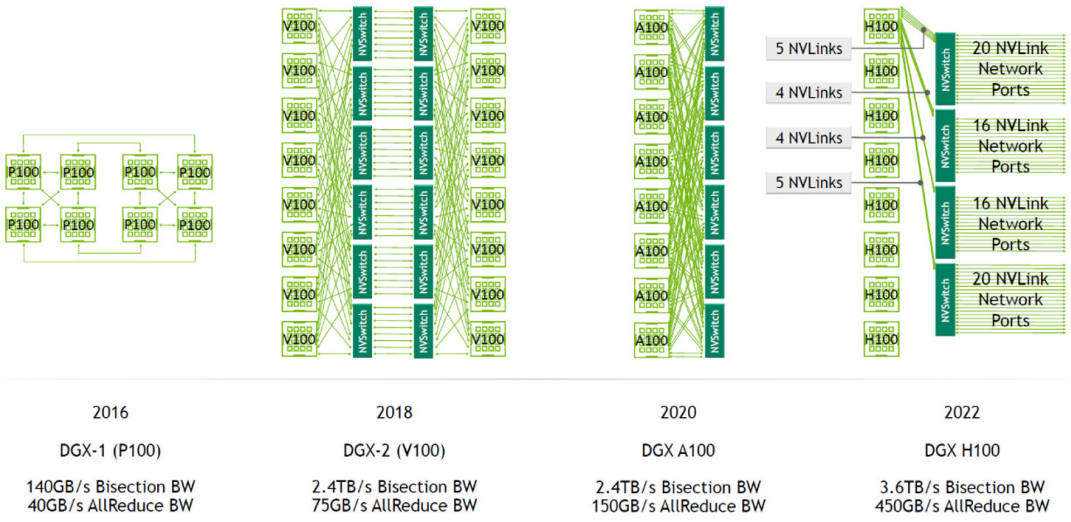


图 3-4 NVLink 卡间互联演进路线

开放的互联标准来源于 OCP 组织发起的开放加速器基础设施项目（Open Accelerator Infrastructure，OAI）[9]，其定义了业界通用的 AI 扣卡模组形态（OCP Accelerator Module，OAM）和基板拓扑结构（Universal Baseboard，UBB），从而降低整机厂家集成多家 AI 芯片的适配难度。基于该标准，目前可实现 128GB/s 卡间互联 P2P 带宽，若采用全互联的拓扑结构，整板 8 卡的聚合带宽可高达 896GB/s。当前主流拓扑为 cube 立方和全互联，未来将增加 Switch 拓扑设计，使卡间 P2P 带宽能力大幅升级（如图 3-5 所示）。在通信协议方面，OAM 推荐采用标准的 PCIe PHY 接口，未对链路层、事务层通信协议进行规范，因此，各 AI 芯片厂家多采用自研的通信协议，如寒武纪的 MLU-LINK、燧原的 GCU-LARE 和壁仞的 B-LINK 等。OAM 和 UBB 的技术生态已日趋成熟，国内外已有整机厂家根据 OAM UBB 规范研发相关服务器，并与多家 AI 芯片开发适配，在国内多地智算中心也有应用落地。

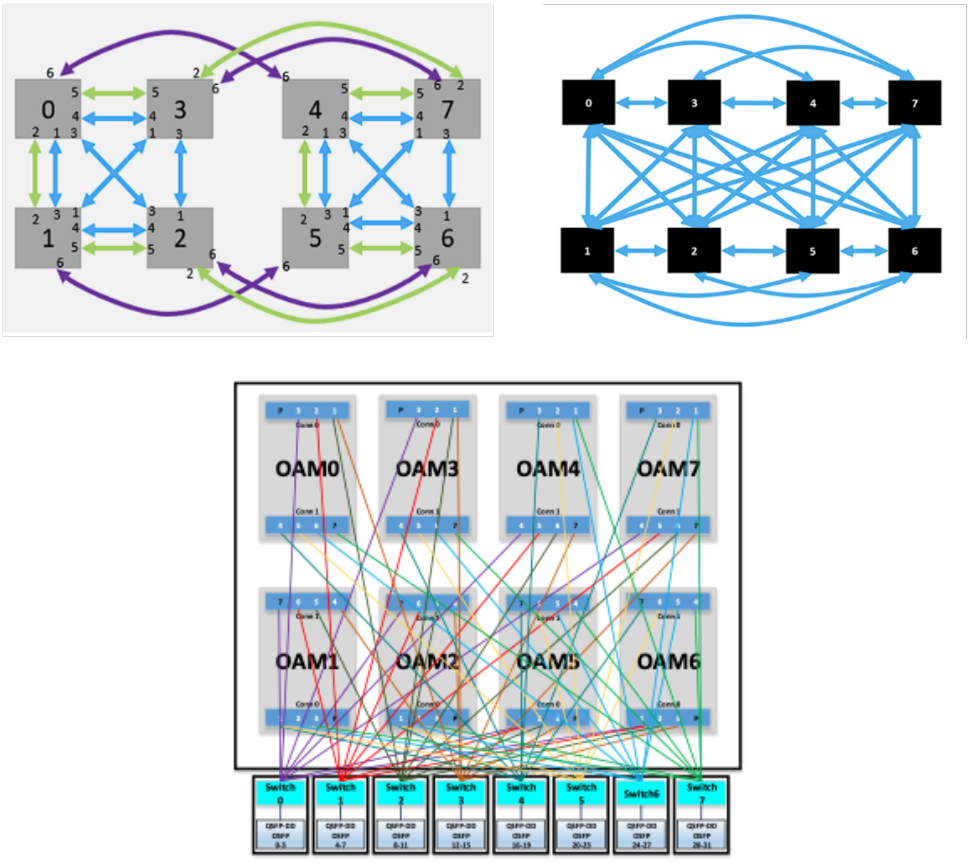


图 3-5 OAM 模块的主流互联拓扑

卡间互联能力与 AI 芯片的吞吐性能、芯片间的互联拓扑以及通信协议设计息息相关。其中，AI 芯片的吞吐性能主要由 SerDes 接口速率和通道数（lane）决定，两者增加会带来互联带宽的提升，但也会引起功耗上升、PCB 布局布线困难等问题，是芯片工程实现的经典 PPA

（Power、Performance、Area）问题；芯片间的互联拓扑决定了整个集群的吞吐能力和扩展能力，当前国际主流水平已经采用 Switch 交换或全互联的拓扑结构达到 8 卡或百卡级别的互联，国内则大多采用 cube 类拓扑实现 8 卡成环连接，相比之下在集群总吞吐和规模能力上均有代际差；芯片的通信协议设计决定了集群的互联效率，同时反过来影响芯片的 IO 设计与卡间拓扑，当前互联协议栈多为 AI 芯片企业自研设计。

3.1.3 未来万亿级模型的卡间高速互联演进建议

基于 Transformer 的大模型演进趋势遵循 Scaling Law[10]，参数量走向万亿级是可预见的必然趋势。新型的算法结构带来了新的分布式训练策略，如专家系统（Mixture-Of-Experts，MoE）并行，**高速通信需求进一步扩展至百卡级别，卡间互联的最优解指向 Switch 交换拓扑，**构建基于交换拓扑的“超级服务器”是未来 AI 基础设施的趋势。目前由于 AI 芯片的互联协议均各自为“栈”，且多数企业缺乏从 AI 芯片到交换芯片的全产品设计能力，导致交换芯片与 AI 芯片之间的互联技术难以匹配，因此交换拓扑的集群方案实现面临强生态门槛，导致芯片互联规模发展受限，在一定程度上制约了 AI 基础设施的先进性。**为降低设计难度，我们建议从统一高速互联协议入手，以实现百卡规模互联为设计目标，收敛技术路线，推动国内高速互联技术生态从能用到好用的跃变。**

目前国内主流方案中，大多数跨机互联主要通过网络协议实现。考虑万亿参数模型对卡间互联的扩展性及开放性要求，可采用**统一的计算总线协议**作为百卡互联的通信方式，逐步推动总线交换芯片的统一，但现有计算总线的设计仍需在带宽、可靠性等方面进行优化：

第一，推动 GPU、AI 加速卡支持统一高效计算总线协议。在大规模并行计算中，各个设备之间高效的数据传输是数据一致性的基本保障，避免由此带来的延迟影响模型训练的效率。统一的计算总线协议避免了不同协议之间的转换，可以确保设备之间数据及时共享。该总线协议的设计应聚焦多个 GPU、AI 加速卡之间在大带宽、低时延的基础诉求，并实现缓存一致性的数据访问，确保简化上层应用研发难度，提升流量控制、拥塞控制、网络无损、重传等通信和数据传输能力。

第二，推动 GPU、AI 加速卡与 CPU、内存等其它核心部件形成开放协议生态。传统的计算架构在解决异构设备互联问题时会使用不同的通信协议和数据格式，协议转换会引入额外的复杂性和延迟，对整体性能产生不利影响。因此，构建多异构设备之间的高速连接通道，将 CPU、GPU、AI 加速卡、DPU、内存、FPGA、SSD 等核心部件进行统一协议互联，使 CPU cache、GPU HBM（High Bandwidth Memory 高带宽内存）、DPU cache、主机 Memory

等设备间进行统一寻址，将有利于降低用户开发难度，提升设备间的系统资源共享（内存和带宽）能力。

第三，推动 GPU、AI 加速卡在功耗和面积上进一步实现集约化设计，满足单芯片计算能力提升和数据中心节能要求。通过引入更高速率的 SerDes IP，对计算总线协议进行优化，减少芯片上所需的硬件资源和物理面积，以减少通信过程中的能量消耗。低功耗的协议有利于降低单芯片能耗，从而提升大规模并行计算的能效。

未来，期望结合计算总线协议推广、产品规模研发、生态系统建设、优化软件和算法与产业开展广泛合作，构建一个灵活强大的计算总线互联生态系统。

3.2 集群间的高速无损网络

3.2.1 InfiniBand 与 RoCE 是当前主流方案

新型智算中心网络从逻辑上可以分为：出口网络、管理网络、参数网络、存储网络和业务网络，如图 3-6 所示。

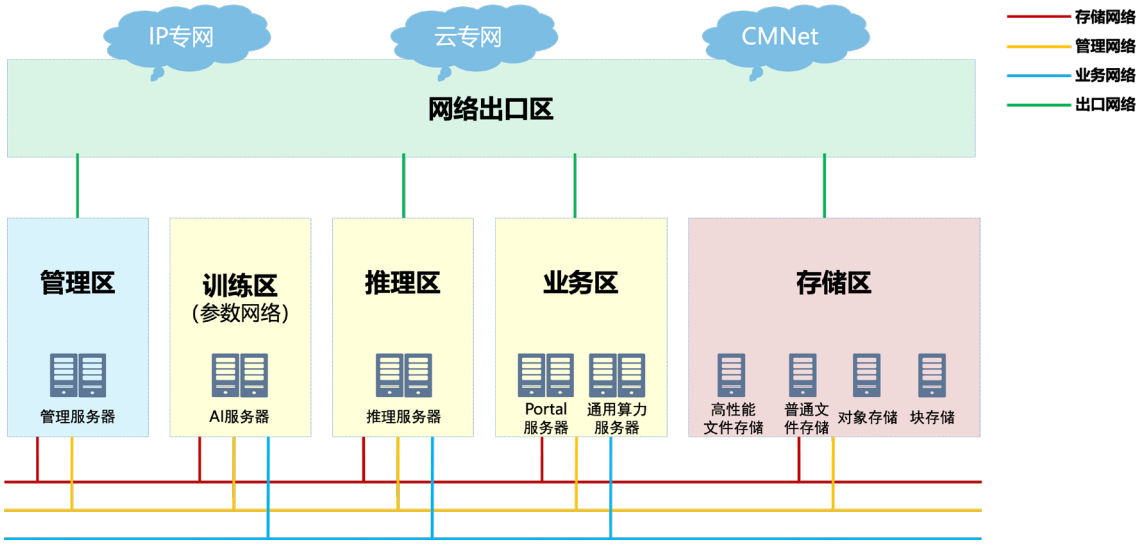


图 3-6 新型智算中心功能模块

其中，参数网络主要用于承载 AI 模型训练业务，其通信流量主要具备周期性、流量大、同步突发等特点。尤其在大模型训练过程中，通信具有非常强的周期性，且每轮迭代的通信模式保持一致。在每一轮的迭代过程中，不同节点间的流量保持同步，同时流量以 on-off 的模式突发式传输，以上通信流量的特点要求参数网络必须具备零丢包、大带宽、低时延、高

可靠等特征。参数网络性能的好坏决定了智算中心提供算力的效率。

现阶段，参数网络存在两种主流的 RDMA 技术，分别是 InfiniBand（简称 IB）和基于以太技术的 RoCE（RDMA over Converged Ethernet），如图 3-7 所示。

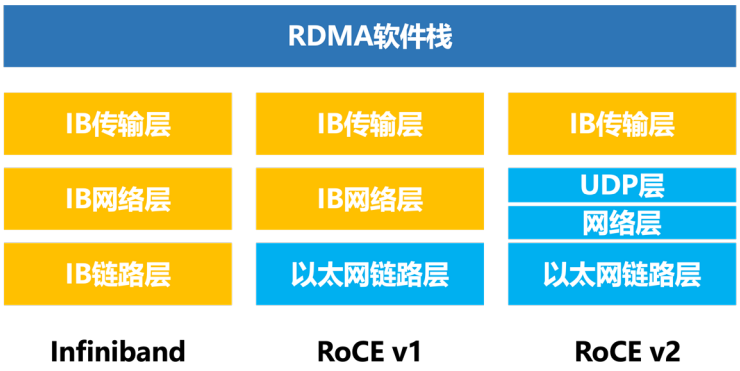


图 3-7 InfiniBand 与 RoCE 协议栈

InfiniBand 由 IBTA（InfiniBand Trade Association）组织于 1999 年提出，是最早出现的 RDMA 技术。InfiniBand 不仅基于网卡硬件实现 L1~L4 层网络协议栈，而且基于集中管理器及端到端的流控机制实现网络无损转发。因此，InfiniBand 机制能够提供超低延迟和超大带宽的网络效果。目前市场上只有 NVIDIA 可提供 IB 交换机、IB 网卡、子网管理器的整套解决方案，但设备采购和维护成本相对较高。

由于从以太网切换到 InfiniBand 网络的成本过于高昂，为推动 RDMA 技术普及，IBTA 在 2010 年提出 RoCE 协议标准，允许应用通过以太网实现远程内存访问，使用者只需要更换网卡，而不需要更换现有的以太网网络设备及线缆就可以享受到 RDMA 带来的网络性能提升和 CPU 负载下降等收益，大幅降低硬件成本和维护成本。

随着智能计算业务的快速发展和部署规模不断扩大，采用 RoCE 技术的智算中心网络在性能和规模方面存在的弊端也渐渐显露出来，主要挑战表现为如下几个方面：

挑战一：传统基于流的等价多路径路由（Equal Cost Multi Path，ECMP）负载均衡技术在流量数小、单流流量大的情况下可能失效，导致链路负载不均。当某些物理链路负载过大时，容易出现拥塞甚至网络丢包。

挑战二：分布式训练的多对一通信模型产生大量 In-cast 流量，造成设备内部队列缓存的瞬时突发而导致拥塞甚至丢包，造成应用时延的增加和吞吐的下降。PFC（Priority-based

Flow Control）和 ECN（Explicit Congestion Notification）都是拥塞产生后，再进行干预的被动拥塞控制机制，它们无法从根本上避免拥塞。

挑战三：业界通过 CLOS 架构搭建大规模分布式转发结构来满足日益增长的转发规模需求，在该架构下，各节点分布式运行和自我决策转发路径导致无法完全感知全局信息和实现最优的整网性能。

3.2.2 全调度以太网突破无损以太性能瓶颈

综合当前所面临的挑战，新型智算中心网络将向三个方向进行演进：一是从“流”分发到“包”分发演进，通过提供逐报文容器动态负载均衡机制，实现单流多路径负载分担，提升有效带宽，降低长尾时延。二是从“推”流机制到“拉”流机制演进，即从被动拥塞控制，到依赖“授权请求”和“响应机制”的主动流控，最大限度避免网络拥塞的产生。三是从“局部”决策到“全局”调度演进，即全局视野的转发调度机制，实现集中式管理运维、分布式控制转发，优化网络性能。

基于以上三大演进方向，**中国移动创新提出全调度以太网（Global Scheduled Ethernet，GSE）技术方案 [11]，打造无阻塞、高带宽、低时延、自动化的新型智算中心网络，助力 AIGC 等高性能业务快速发展**（如图 3-8）。

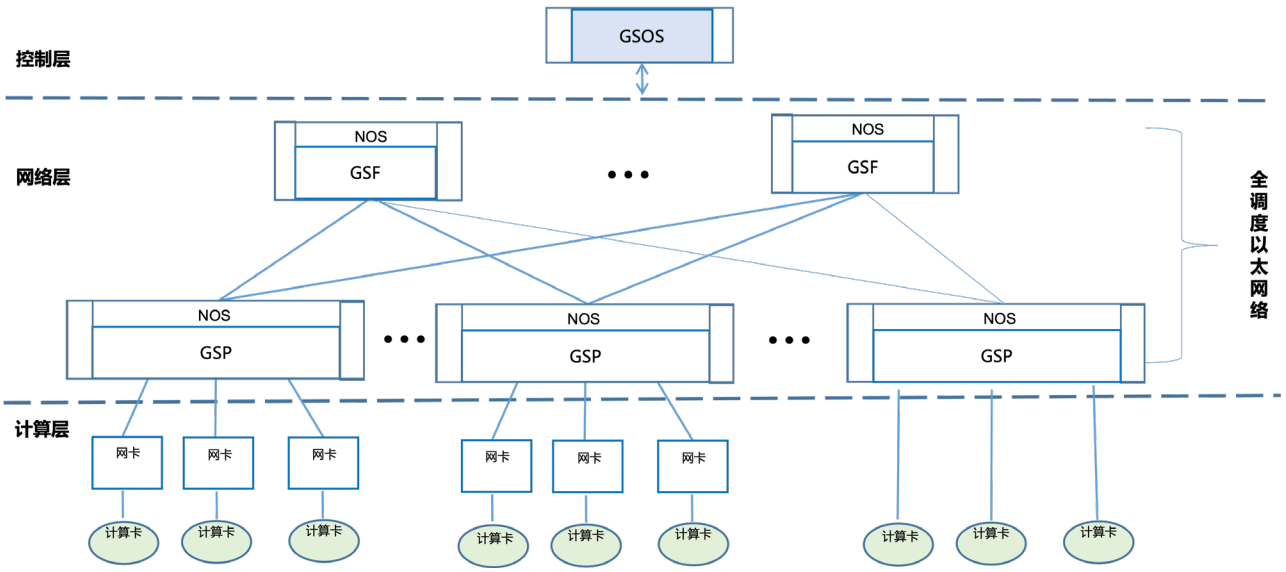


图 3-8 全调度以太网（GSE）技术架构

如图 3-8 所示，全调度以太网（GSE）技术架构主要包括计算层、网络层和控制层 3 个层级，包含计算节点、网络边缘处理节点（Global Scheduled Processor, GSP）、网络核心交

换节点（Global Scheduled Fabric, GSF）及全调度操作系统（Global Scheduled Operating System, GSOS）4 类设备。

计算层为 GSE 网络服务层，包含高性能计算卡（GPU 或 CPU）及网卡；GSE 网络层主要实现 GSP 和 GSF 协同，实现基于报文容器的转发及多路径负载、基于报文容器的全局视野的流量调度等技术融合的交换网络；控制层主要包含全局集中式 GSOS，以及 GSP 和 GSF 设备端分布式 NOS（Node OS），实现集中式管理运维及分布式控制转发。

计算节点即服务器侧的计算卡、网卡，提供高性能计算能力。GSP 即网络边缘处理节点，用以接入计算流量，并对流量做全局调度；流量上行时具备动态负载均衡能力，流量下行时具备流量排序能力。GSF 即网络核心交换节点，作为 GSP 的上一层级设备，用于灵活扩展网络规模，具备动态负载均衡能力，以及反压信息发布能力。GSOS 即全调度操作系统，提供整网管控的集中式网络操作系统能力。

综合考虑分布式 NOS、集中式 SDN 控制器的优势，全调度以太网的 GSOS 分为全调度控制器、设备侧 NOS 两大部分。如图 3-9 所示，GSOS 维护全局网络信息，实现 DGSQ（Dynamic Global Scheduling Queue）系统的建立和维护。NOS 运行设备自身网络功能，提升系统可靠性，降低部署难度，全面提升 GSE 网络自动化及可视化能力。

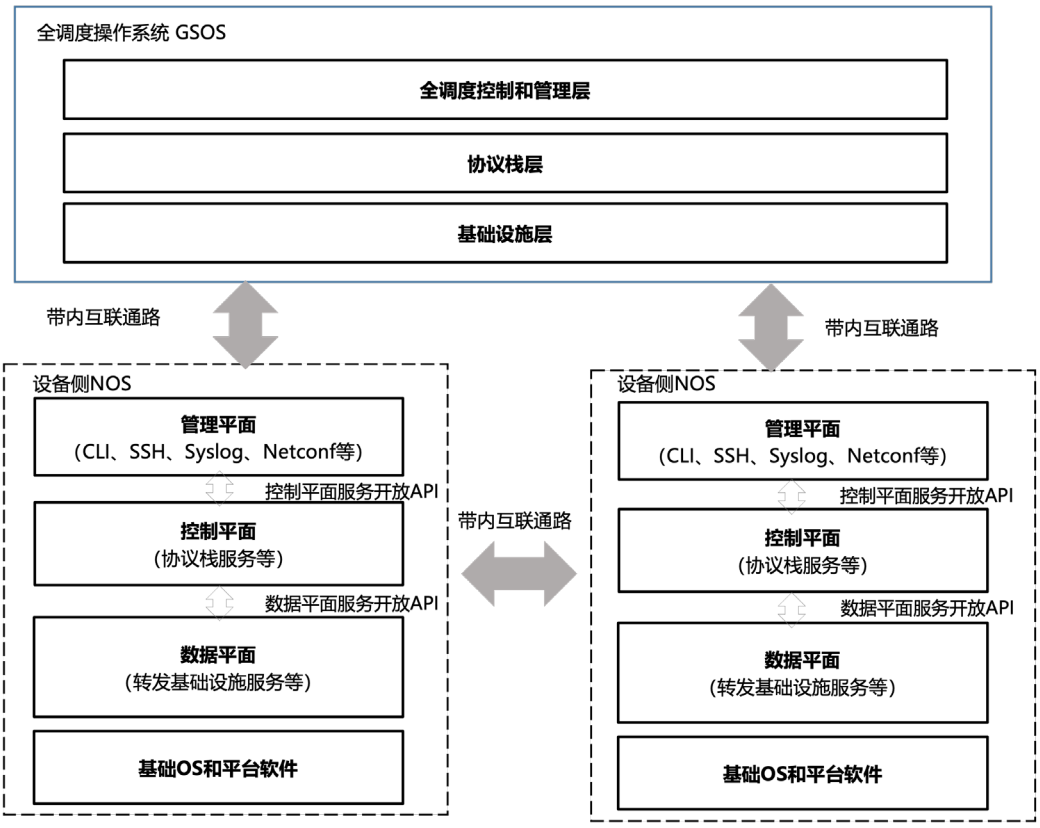


图 3-9 全调度以太网操作系统架构

智算中心网络通常采用胖树（Fat-Tree）架构，任意出入端口之间存在多条等价转发路径，智算业务流量存在“数量少，单流大”的特点，传统以太网逐流负载分担方式导致链路利用率不均，从而引起网络拥塞。单流多路径是提升智算中心网络有效带宽、避免网络拥塞的关键技术手段。GSE 技术架构提出一种基于报文容器（Packet Container，PKTC）的转发及负载分担机制，即根据最终设备或设备出端口，将数据包逻辑分组，并组装成长度较长的“定长”容器进行转发，属于同一个报文容器的数据包标记相同的容器标识，沿着相同路径进行转发，以保证属于同一个报文容器的数据包能够保序传输，如图 3-10 所示。

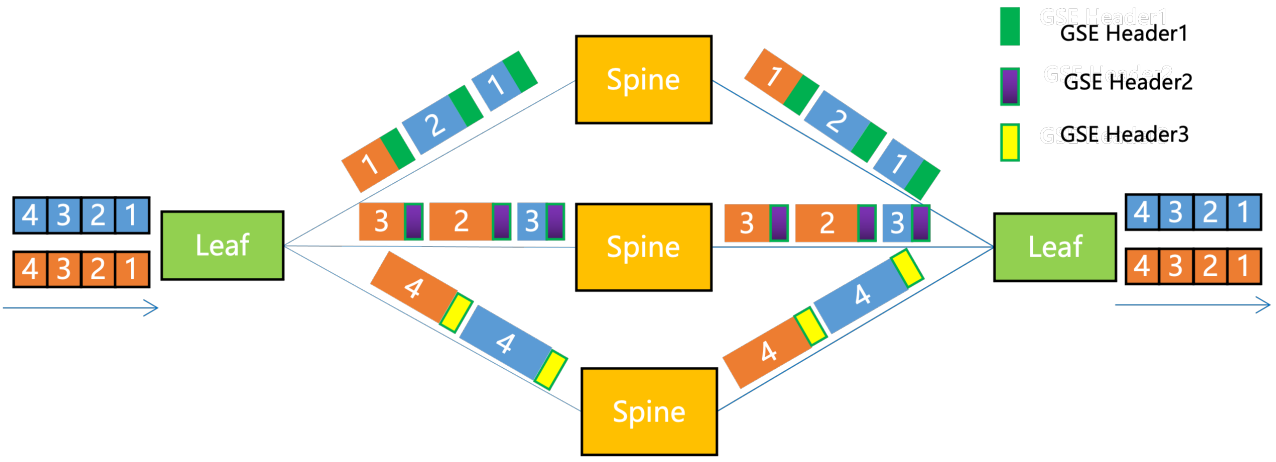


图 3-10 报文容器转发示意图

在负载均衡调度时，报文容器被作为传输单元，但由于报文是逻辑组装，无需额外的硬件开销来对数据包进行组装和还原。在网络中转发时添加的报文容器标识，仍以数据包的形式传输，且无冗余数据填充的问题，带宽损耗小。

另一方面，由于模型训练流量的特殊性，网络会出现“多打一”的流量，引发网络拥塞。如图 3-11 所示，GSP1 的 A1 口和 GSP3 的 A3 口同时向 GSP2 的 A2 口发送流量，且流量相加大于 A2 的出口带宽，造成 A2 口出口队列拥塞。这种情况仅通过负载均衡是无法规避的，需要全局控制保证送到 A2 的流量不超过其出口带宽。

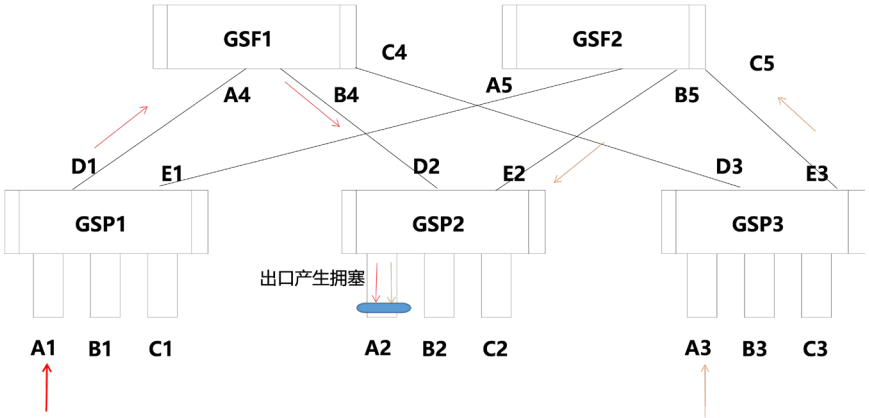


图 3-11 网络 Incast 流量发生场景

基于 DGSQ 的全局调度技术如图 3-12 所示，在 GSP 上建立网络中所有设备出口的虚拟队列，用以实现本 GSP 节点到对应所有出端口的流量调度。本 GSP 节点的 DGSQ 调度带宽依赖授权请求和响应机制，由最终的设备出口、途经的设备统一进行全网端到端授权，保证全网中前往任何一个端口的流量既不会超过该端口的负载能力，也不会超出中间任一网络节点的转发能力，可降低网络拥塞发生的概率，减少内部反压机制的产生，提升网络性能。

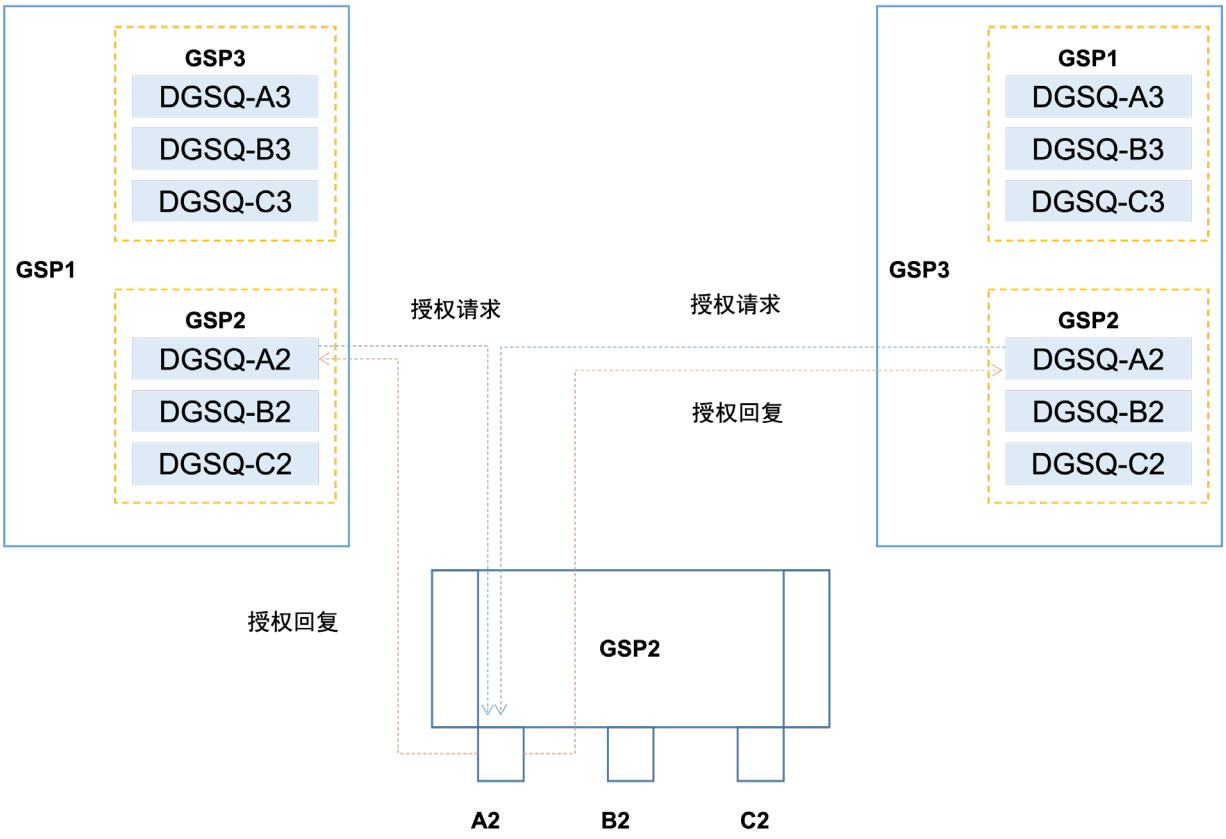


图 3-12 基于 DGSQ 调度流程

作为一种标准开放的新型以太网技术，GSE 可采用网卡侧无感知的组网方案，即网络侧采用 GSE 技术方案，网卡侧仍采用传统 RoCE 网卡。此外，也可以结合网卡能力演进，将 GSE 方案各组件的功能在网络组件中重新分工，将部分或全部网络功能下沉到网卡侧实现。也就是说，在未来的实际应用中，可以将 GSP 的功能全部下沉到网卡以提供端到端的方案，也可以将网络的起终点分别落在网络设备和网卡上，为后续网络建设和设备选型提供灵活的可选方案。

3.2.3 智算中心网络关键技术演进

1）速率体系升级，功耗成为挑战：在交换芯片方面，交换芯片作为网络设备的心脏，直接决定设备能力。当前业界单芯片容量已达 51.2T，SerDes 速率也演进到 100G。在接口带宽方面，传统通用数据中心服务器接入带宽以 10G、25G 为主，而 AI 集群普遍采用单卡 100G/200G 的高性能网卡，最高端网卡已经达到 400G 接口。而网络接入层交换机需配套 100G/200G 甚至更高速率的 400G 交换机，汇聚层交换机端口也演进到了 800G。因此，芯片容量、单通道数据传输速率的大幅提升和对光模块速率、数量要求的提高，使得网络设备本身功耗极速升高，网络耗能占比不断提升。在此背景下，CPO（Co-packaged Optics）和 LPO（Linear-drive Plugable Optics） 等技术陆续被提出，其中 CPO 技术将硅光模块和 CMOS（Complementary Metal Oxide Semiconductor）芯片集成，缩短芯片和模块间的走线距离，降低成本和功耗。该技术虽前景可观，但对现有网络建设和运维体系等方面带来许多新挑战。LPO 技术将传统光模块内部 DSP 功能集成到交换芯片中，降低光模块层面信号处理的功耗和延迟，但由于对 SerDes 以及模块侧光芯片要求较高 ，技术可行性及产业成熟度仍待验证。

2）低时延转发，FEC 技术是关键：随着端口速率的不断提升，高速信号完整性的挑战也越来越大，需要不断引入更为强大的 FEC（Forward Error Correction，前向纠错）算法。FEC 越强大其编解码复杂度也越高，所增加的时延也越大，100G 以上的速率 FEC 所占用的时延已经达到整体转发时延的 20% 左右。

FEC 的过程又可以分为检错逻辑和纠错逻辑。在低速的 FEC 处理中往往没有做上述流程的区分，但随着速率提升、检测及纠错逻辑的复杂，细分差异化处理会变得越来越有意义。检错和纠错分离技术可提前校验数据块内是否存在误码。在无错情况下，可旁路 FEC 译码流程，消除无错场景下 FEC 收帧和译码时延，降低无错情况下的接口时延，消除高增益 FEC 码字的时延弊端；有错的情况下，会进一步进行纠错处理。因为发生误码的概率远小于无误码，

所以此方式可以优化端口的平均转发时延。灵活 FEC 技术可以根据链路的误码率状态，自动选择合适的 FEC 纠错算法，以便在保持可靠性的同时提供低延迟。

3）高安全防护，物理层加密有优势：随着生成式 AI 等应用的发展，对海量算力芯片间高吞吐、低时延数据传输的需求更为迫切，这些数据不仅涉及用户隐私，也关系到企业的安全生产。为了应对日益严峻的数据安全挑战，要对以太网传输链路提供数据安全加密能力并关注数据加解密带来的时延与开销。目前以太网已部署的存量设备可能存在硬件芯片无法更换的情况，链路级数据加密技术需要在现有网络设备的基础上具备前向兼容能力。

现有 MACSec 等网络安全加密技术难以完全覆盖链路层及以上协议层的安全加密。如基于优先级的流量控制帧无法加密帧头部以及掩盖帧发送频率、帧长等流量特征，该方式难以有效防止流量分析攻击，存在安全漏洞。PHYSec 技术将物理层加密的理念与以太网物理层技术相融合，以实现低开销、低时延、高安全和协议透明等特性的安全加密机制，满足数据链路层及所有上层协议的信息防护。

4）拥塞控制，端网协同是核心：由于网络中流量的随机性以及路径的多样性，拥塞的出现不可避免。网络出现拥塞后，会造成排队时延增大、网络利用率降低等影响，导致应用性能出现恶化。传统的拥塞控制以被动拥塞控制为主，即收到拥塞信号后被动探测式地调整速率。典型的如 DCQCN（Data Center Quantized Congestion Notification）算法，发送端根据接收到的 ECN 标记报文，利用 AI/MD 机制（Additive-Increase/Multiplicative-Decrease，线性增速乘性降速）调整发送速率。由于 1 个比特的 ECN 信号无法定量地表示拥塞程度，发送端设备只能探测式地调整发送速率，导致收敛速度慢，性能较差。

目前，业界典型的优化思路分为两类：第一类是更加精细化的被动控制，如 HPCC（High Precision Congestion Control，高精度拥塞控制），该算法利用相比 ECN 更精细的信息，提高调速的准确率，避免长时试探；第二类是提前预留 / 主动分配式的主动控制，如 HOMA（一种接收端拥塞控制算法）等，主动为后面的包做资源预留以及分配，避免拥塞的发生。

4 新算效——重塑计算架构

4.1 下一代 AI 芯片设计思路

以 GPU 为代表的高性能并行计算芯片架构和以针对 AI 领域专用加速（DSA, Domain Specific Architecture，DSA）为代表的芯片架构是目前两大主流 AI 芯片设计思路。GPU 设计初衷是为了接替 CPU 进行图形渲染，图形处理涉及到相当多的重复计算量，因此 GPU 芯片上排布了数以千计的，专为同时处理多重任务而设计的小计算核心。随着 AI 深度学习算法的逐渐成熟，GPU 芯片开始引入 AI Core/Tensor Core 等电路来实现矩阵乘运算的加速。因此，GPU 比 CPU 拥有更强的大规模并行计算和浮点运算能力。不同于 GPU，AI DSA 芯片是一种针对神经网络计算的专用处理器，主要功能是加速神经网络的数据处理、传递和反向传播等操作，因为芯片架构是专用设计，相比 GPU，减少了深度学习不常用的 FP64 等标量计算单元。AI DSA 芯片在功耗、可靠性、芯片体积、性能等方面都有巨大的优势，但由于电路设计是定制思路，芯片开发周期较长，在通用性和可编程性方面相比 GPU 架构较弱，当前技术和生态还处于多而不强的局面。 GPU 生态代表的是英伟达，AI DSA 是 Google

TPU、华为昇腾、寒武纪思元等。

面向未来万亿模型的兴起到大模型应用逐渐落地，算法需要挖掘海量数据进行计算，无论是 GPU 还是 DSA 芯片架构设计，均面临性能瓶颈。一是内存带宽的制约，已经成为整个系统的性能瓶颈。二是海量内存和计算单元间的频繁访问切换，导致整体功耗激增。其次，大模型技术的快速迭代，硬件如何适配算法也仍是难题。针对下一代芯片设计思路，有以下方向:

一是存算一体化设计思路，解决存储带宽和访存功耗的问题。未来 10 年是计算架构变革的新十年。计算存储一体化已经是业内一大研究方向。存算一体芯片将计算单元和存储单元合一，使得芯片在提升计算和数据吞吐量的同时显著降低功耗。

二是引入稀疏化计算能力解决大量运算带来的功耗问题。虽然千亿、万亿模型相继提出，但并不是每个神经元都能有效激活，这个时候稀疏计算可以减少无用能效。在推荐和图神经网络的场景，稀疏已成为常态。

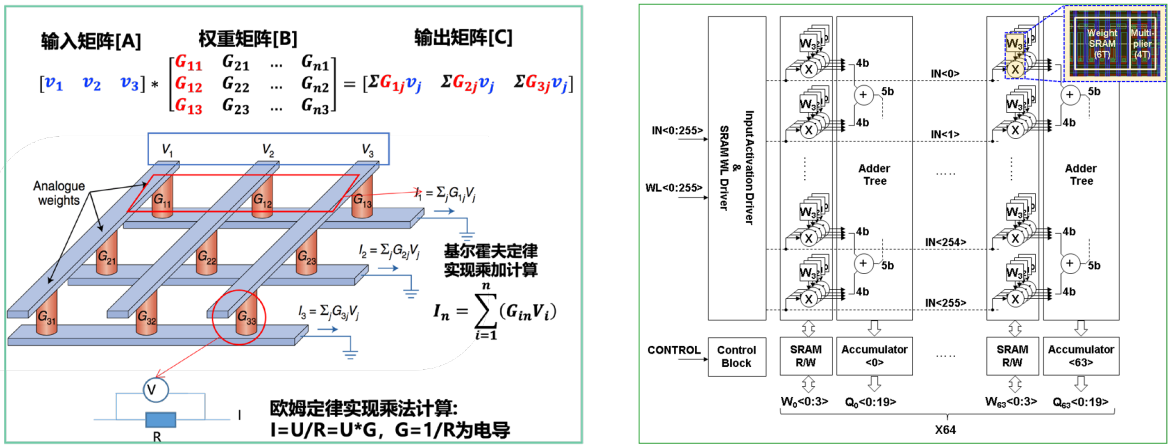
三是在芯片设计上支持更加复杂的 AI 算子。如 Transformer 结构或者是在 NLP 和语音领域经常用到的动态 shape，需要合理地分解、映射这些不同复杂结构的算子到有效的硬件上，算法和芯片协同设计是下一代 AI 芯片设计的重要思想之一。

四是芯片需要支持更低的推理时延。为加速预训练大模型的应用落地，产业已开始研究通过量化、蒸馏、剪枝等手段来使大模型小型化。大模型推理场景下，计算负载出现了混合精度，8bit 甚至更低精度，如何提升芯片推理的实时性和并行能力也将是一大研究方向。

4.2 存算一体构建新型计算范式

存算一体作为新型计算范式，基于在存储原位实现计算的本质，打破了冯诺依曼存算分离架构，避免了频繁的数据访问和搬运带来的功耗激增的问题，大大缓解了 AI 芯片性能提升的瓶颈。同时，由于新型智算中心承载的 CNN、Transformer 等主流模型架构，矩阵乘加运算占据了大量算力（Transformer 中 45-60%，CNN 中 90% 以上的运算均为矩阵乘加），存算一体的架构成为高效完成矩阵乘加的重要选择。存算一体可通过 RRAM、SRAM、MRAM、Nor Flash 等介质实现，多介质共存可以发挥不同介质在成熟度、读写次数等方面的优势 [12]。

存算一体通过模拟计算或数字计算或二者相结合的方式提供存算能力，如图 3-13 所示：



a）模拟存算一体　　b）数字存算一体

图 3-13 模拟和数字存算一体

存算一体在 NICC 的主要应用是大模型推理。考虑到不同的模型结构，存算一体充分利用非规则稀疏性，以达到与存算阵列的最佳适配，并实现能效最大化。以复旦大学 ISSCC 2023 发布的论文为例 [13]，其应用了基于蝶形数据分配网络的稀疏前馈计算架构（如图 3-14），结合对应的存内阵列设计和电路实现，能够在 28nm 工艺下，达到现有 Transformer 加速器 3.2 倍至 9.7 倍的能效。

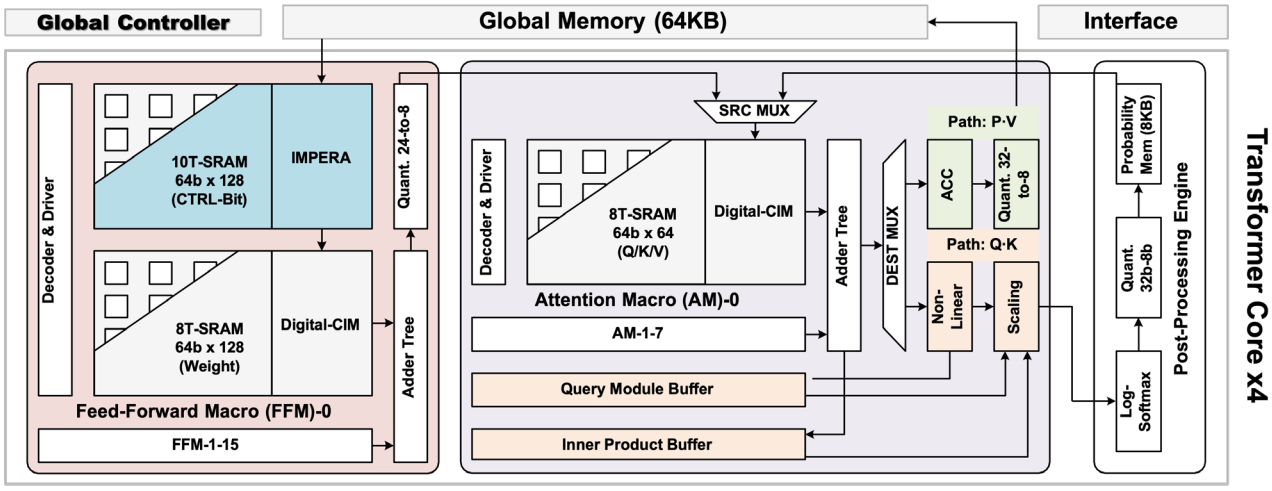


图 3-14 Transformer 加速器的存算一体实现示意

当前，面向智算中心多核、多芯片的存算一体架构方案将成为未来存算一体研究和商用落地的重点方向。在此场景下，有以下三点问题需重点解决：

1) 如何与算法结构协同：通过改进的存算一体阵列架构更好的适配稀疏 Transformer，使用分块结构化稀疏、动态激活值稀疏以及特定 Transformer 稀疏等方式，选择参与计算的存算单元，并结合定制的加法树电路减小面积，提升计算能效，从而提升推理效率。

2) 如何与精度需求协同：通过可变精度存算一体阵列架构更好适配大模型推理的精度需求，使用数字存内计算和模拟存内计算混合、双生多 bit 等方法，实现 INT8 和 BF16 等混合精度计算。

3）如何与封装能力协同：通过 Chiplet 技术同时满足存算一体专用高性能、通用可扩展要求，提升算力和 IO 带宽，减少访存瓶颈; 通过 3D 堆叠等封装技术将存内计算（CIM）与近存（PNM）和存内处理（PIM）技术结合，为访存密集型应用提供大容量高带宽的计算能力。

现阶段的存算一体芯片在介质优化、集成规模、工具链支持、算法适配、产业生态等方面还面临诸多挑战，导致应用普及较慢，建议锚定智算核心应用，推出样板产品，突破上述关键挑战，在成熟工艺实现性能反超。

4.3 DPU 实现计算、存储和网络的深度协同

DPU 作为 CPU、GPU 之后的数据中心第三颗大芯片，本质是围绕数据处理提供网络、存储、安全、管理等基础设施虚拟化能力的专用处理器。面对智算业务场景，中大规模模型训练和推理任务对网络和存储 I/O 的时延提出了更极致的性能需求，**DPU 可在智算领域解决三大关键问题，与计算、网络、存储深度协同，**助力算效提升。

1) 统一云化管理：智算服务场景存在裸金属、容器、 虚机多种方式部署需求，如何实现 AI 节点并池管理提高计算资源利用效率，成为关键的业务痛点，DPU 是最佳的解决方案。通过 DPU 可提供计算资源快速发放和回收等底层支撑能力，使弹性裸金属特性和虚机一致，支持云盘启动，完成灵活的存储分配，实现存储多租户隔离并缩短容灾时间，交付效率提高 10 倍。

2) 高性能存储卸载及加速：大模型训练推理业务的模型本身以及训推所需的数据需要 PB 级储存，本地存储性价比低，远端存储集群成为最优选择。分布式存储设备面对上千计算节点，需要满足多用户并行使用时产生的海量数据读取及加速数据收敛需求，单节点存储带宽叠加

后对存储系统提出更高的性能要求。DPU 产品可以提供专用的高速存储单元来处理和管理大量的数据，提供高带宽和低延迟的存储访问，实现 NVMe-OF 存储加速，同时可配合训练框架进行文件系统卸载，实现训练数据格式统一化，实现不同来源的数据接入，进一步加速训练和推理过程。

3)RDMA 网络协同优化：智算集群由大量的智算服务器节点组成分布式系统，节点间通信基于 RDMA 低延迟 Fabric 网络进行连接，可通过 DPU 产品提供 GPUDirect RDMA 能力及 RDMA 大规模队列资源增强能力，借助 DPU 优秀可编程特性，协同网络侧进行高性能 RDMA 及无损网络优化，实现高吞吐低延迟的网络能力。最终解决大规模并行训练场景多机间高速互联问题，提升网络传输效率，构建端网协同的高性能智算架构。

为解决上述关键问题，新引入的 DPU 部件作为智算服务器的 IO 入口需对原有的网卡部件进行替代，硬件典型配置上可能出现两种替代模式（如图 3-15）：

1)DPU 替代服务器中存储面及管理面网卡，**作用在 CPU 域，**原配置中多块 CPU 下挂的多块网卡被一张 DPU 卡替代。此配置用于解决云化平台统一、高性能存储加速需求。

2)DPU 替代服务器中参数面网卡，**作用在 GPU 域，**原 PCIeSwitch 下挂的多块大带宽 RDMA 参数面网卡被 1:1 替换为 DPU 卡。此配置用于解决 RDMA 网络协同优化需求。

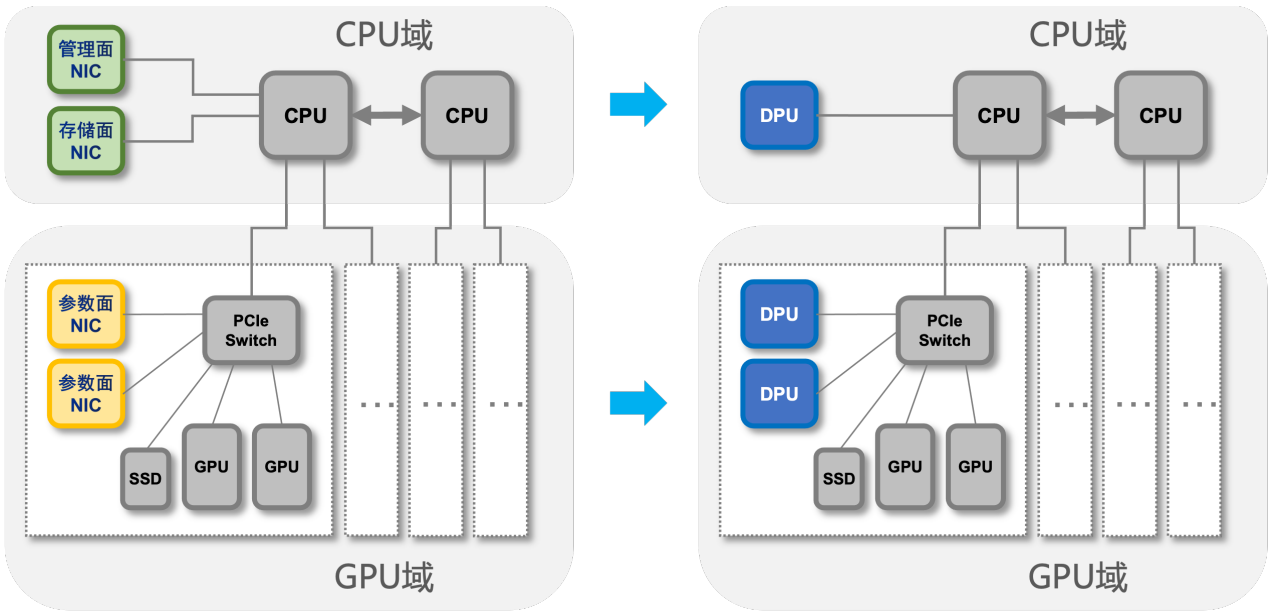


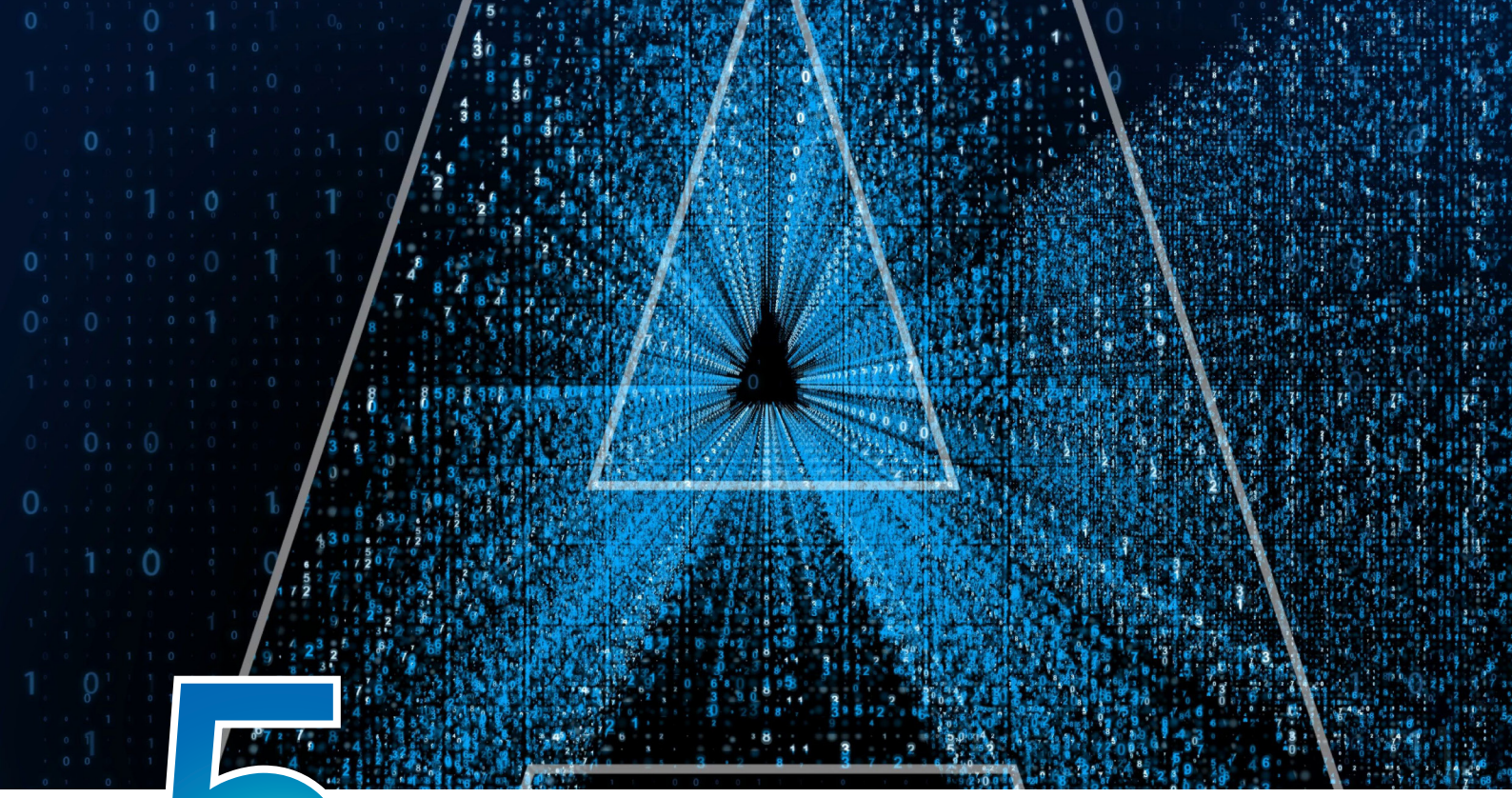
图 3-15 智算中心引入 DPU 两种模式

面对 DPU 在智算场景的试验试点及规模应用，当前仍面临三大核心挑战：

在云平台侧，DPU 软硬融合层的标准化是制约 DPU 通用化的主要问题。DPU 本质是云化、虚拟化技术从软件实现向软硬结合发展的结果，技术架构与云计算关系密切，存在耦合，DPU 虚拟化技术栈在技术迭代中差异化发展，不同产品的同一技术的实现路径多样，软件实现方式差异大。亟需解决业界异厂家 DPU 与云平台软件定向开发适配成本高的问题。建议围绕管理、网络、存储、计算、安全五大软件系统，推动 DPU 软件功能要求和交互接口标准化，并分阶段推进。

在网络侧，网络技术创新需要与 DPU 深度协同。智算业务要求零丢包、低时延、高吞吐的网络能力，RDMA 网络是智算中心高性能网络的首选，头部企业纷纷布局自研 RDMA 协议栈及无损网络相关技术。DPU 作为服务器的 IO 出入口，是网络与存储必经之路，网络技术创新需要与 DPU 深度协同，实现算力无损，助力算效提升。

在硬件侧，亟需优先引导服务器整机层及 DPU 部件层标准化及通用化。重点围绕服务器结构及供电、散热、带外纳管方案、上下电策略四大方向进行统一，为 DPU 与上层软件的深度整合及生态繁荣提供底层支撑。



5

新存储——挖掘数据价值

5.1 计算与存储的交互过程

大模型训练是一项复杂而耗时的任务，类似 GPT-3 级别的模型训练数据集通常很大，无法完全加载到内存中，需要分批次的从外部分布式存储中读取数据并加载到 GPU 的 HBM 上。如图 3-16 所示，从用户上传原始数据集到最终完成模型训练，并对用户提供已训练模型结果，整个过程存在着计算与存储系统密切的数据交互。

1）数据上传：大模型预训练阶段首先需要获取训练数据集，这些来自互联网、书籍、论文的数据需要进行预处理和清洗，包括分词、去除噪声和非常见词汇，以确保训练数据是高质量且可靠的。数据集准备好之后上传到存储系统中。由于对象存储具有普遍的 API 支持，可以提供灵活的数据访问方式，数据集通常会上传到对象存储中。大模型训练的数据集可达 TB 量级，且主要以大文件大 IO 写入为主，存储系统需要保证足够和稳定的吞吐性能。

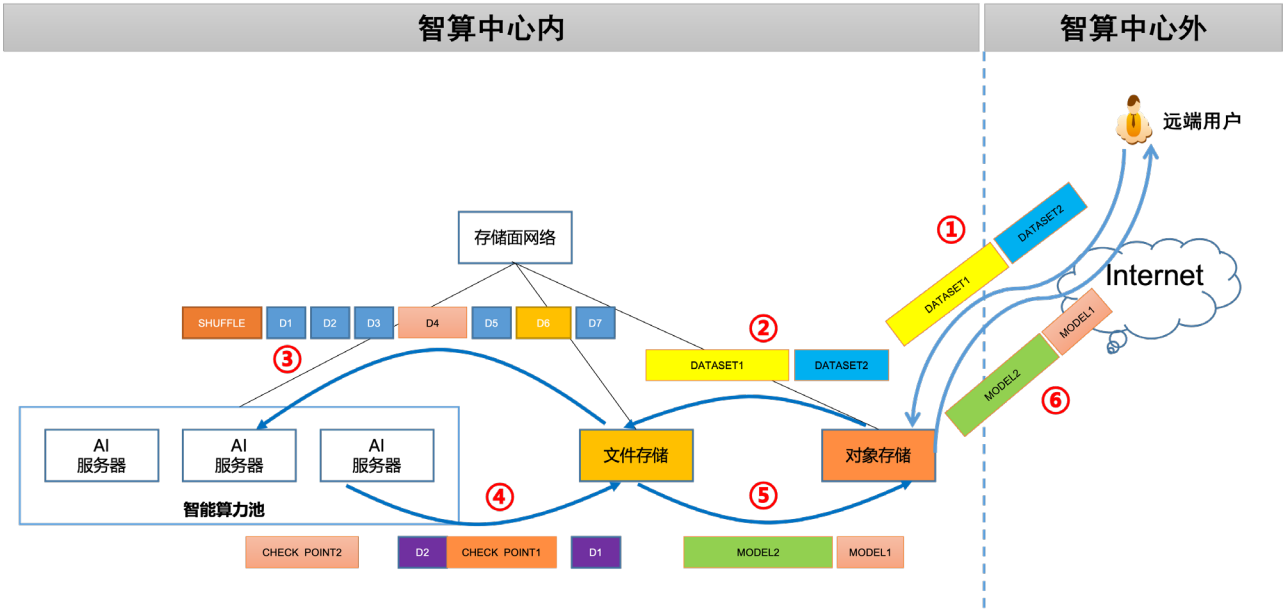


图 3-16 大模型训练计算与存储的交互过程

2）数据转移：由于文件存储具有更高的 IO 性能，对于小文件和随机 IO 有较好的支持，且与 TensorFlow、PyTorch 等训练框架的兼容性更好，适合在训练过程中进行高效的读取和写入操作，因此在模型训练开始之前，需要把数据集从对象存储复制到文件存储中，这个过程中，IO 类型以大文件大 IO 顺序读写为主。

3）数据读取：数据集放入文件存储后，还需要进行进一步预处理。CV 类数据集通常需要先对图片序列化并添加类别标签、图像尺寸等元数据，自然语音类数据集则需要对语音文件进行切分，转换为训练框架实现代码期望的采样率和格式，例如 16K 采样 wav 格式。数据集准备就绪后，模型将基于随机初始化的权重启动训练。整个数据集会被随机打散，称之为 shuffle，然后数据被分成多个小的批次（batch），后续计算节点将以批次为单位从文件存储系统读取数据，并缓存到 GPU 的 HBM 中。

4）归档写回：由于 HBM 是易失性存储，一旦在训练过程中发生意外中断，训练数据将全部丢失，因此基于 Checkpoint 的“断点续训”机制非常关键，我们需要将模型训练过程中的数据周期性地保存到外部持久性存储中，一旦发生中断可以从最后一次保存的参数处重新开始训练，从而节省大量的时间和经济成本。此外，文件存储还用于跟踪记录模型训练过程中的各种指标，包括损失函数的变化、准确率的提升等，以便后续支持可视化的模型训练策略优化分析。保存 checkpoint 和过程文件等操作，主要负载是大文件大 IO 写操作，对文件存储压力不大。

5）模型复制：模型训练完成后，最终的模型权重会被写入到文件存储中保存，用于模型推理或者以 MaaS 的服务模式给外部用户使用。由于对象存储便于对外共享，模型需要从文件存储复制到对象存储上，这个环节 IO 类型以写入大文件为主。

6）模型下载：用户基于自身应用特点，从对象存储下载训练好的模型。

5.2 智算场景存储面临的三大挑战

智算中心存储设施承载海量非结构化数据，根据业务需求通常包括文件存储、对象存储、块存储三种传统类型存储，不同类型的存储池相互分立，各自使用独立命名空间，这种部署方式面临三方面的挑战：

1）存储性能：大模型训练过程中，呈现出大量小文件小 IO 读操作负载特征，是对传统文件存储性能的巨大考验，往往会遭遇“存储墙”性能瓶颈，使 GPU 经常处于等待数据的状态，降低 GPU 利用率，增加训练耗时。目前业界通过升级分布式文件存储节点硬件配置、提高前后端组网带宽、使用 NVMe 全闪存储介质、软件优化小文件处理等手段，提升文件存储性能。

2）存储容量：整个模型训练的过程涉及到文件存储和对象存储之间的大文件大 IO 数据拷贝，由于每种存储内部均设有副本和 EC 冗余机制，以及存储 SSD 硬盘的写放大效应，整个智算中心需要规划超过数据量本身数倍的存储空间，对智算中心的存储容量带来前所未有的挑战，亟需更高效的智算存储解决方案。

3）存储调度：超大规模的模型训练，未来可能需要实现跨地域多中心并行训练，以有效拉通整体的算力和存储能力，这要求存储具有跨地域统一命名空间、统一存储资源调度和足够高的端到端数据交互性能，但是当前仍然受到存储现有技术架构和长距离数据传输性能的限制，业界还没有成熟可行的解决方案。

5.3 多协议融合存储贯通异构数据

文件存储和对象存储虽然架构和协议不同，但都使用元数据机制存储非结构化数据，因此文件存储和对象存储可合并升级为融合存储。目前已经有一些开源或商业的解决方案，例如 MinIO、Ceph RGW 等，这些方案提供统一接口，实现数据交互优化。这些方案提供统一的接口实现两套系统数据交互的优化。使用融合存储，用户只需将数据存储在一个地方，避免

了数据复制和迁移，节省存储空间和管理成本，还可支持多种存储协议的同时访问。通过统一的存储架构，文件、对象等多种协议存储可以更加紧密地协同工作，实现数据的共享和传输。

实现存储协议融合需要关注协议转换语义损失和安全访问策略差异两大问题：

1）协议转换语义损失：文件存储和对象存储采用不同的数据存储范式，一些文件系统独有的特性无法完全映射到对象存储的模型中，例如，文件系统支持硬链接、符号链接等特性，在对象存储中无法直接体现；文件系统的部分访问控制列表（ACLs）可能无法转换到对象存储的访问策略中，因此在具体使用时，会造成关键语义的损失。

2）安全访问策略差异：在文件存储系统中，访问控制通常是基于传统的文件权限，例如 POSIX （Portable Operating System Interface）权限模型。而对象存储通常采用基于角色的访问控制或基于资源的访问策略。因此，将文件存储中的文件通过对象协议访问时，需要适配修订访问策略，可能导致安全性上的差异。

随着智算中心等场景对于存储的容量高效率利用和访问灵活性的需求，协议融合存储已成为发展趋势。对于融合存储语义损失和安全策略差异的挑战，可通过**原生协议融合存储**来解决。原生协议融合指的是，制定统一的存储框架，在底层将元数据和数据实体抽象为元素，根据用户发起的协议要求，将数据元素组合起来提供服务。要实现协议的原生融合，需要产业界：1）共同设计一致的数据模型标准，兼容文件存储和对象存储的特性，最大程度地保留文件系统中的语义和特性，并在对象存储中准确呈现； 2）将文件系统的元数据（如文件权限、访问时间、所有者等）转换为对象存储系统的元数据格式，实现两个系统的元数据准确映射，以便保持一致性; 3) 定制统一的访问控制策略，确保在融合存储中实现相似的访问权限控制; 4）实现 QoS、分级和配额等高级特性共享，实现扩展性、性能、可服务性的能力统一等。

5.4 全局统一存储打破单体局限

跨地域多数据中心之间的全局统一存储，可以实现全局存储资源抽象，形成逻辑上统一的命名空间，一方面使得数据在不同数据中心之间的复制和同步更加透明，确保了数据的一致性和高可用性。另一方面，可以通过负载均衡策略，使得应用程序可以就近访问数据中心，降低访问延迟。这样的设计使得上层智算应用可以在不同数据中心之间无缝地访问和操作数据，

而无需担心数据存放的物理位置，为实现跨域的分布式并行训练奠定数据基础，也使得大模型训练不再受到单体智算中心存储容量的限制。

跨地域全局统一存储当前还属于中远期的技术，处于起步探索阶段，需要产业界重构存储架构设计，并通过制定统一接口标准等方式，解决跨厂商存储资源统一调度的问题。与此同时，跨地域长距离的数据交互时延受到网络传输性能极限的约束，虽然当前可通过负载均衡策略就近选择数据中心，基于性能和距离等因素对数据进行热冷分级存储调度，长期来看，仍有待存储、网络、传输等多个专业领域在基础技术层面取得创新突破。

5.5 基于计算总线构建统一内存池

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、HBM、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，必须通过手动管理数据传输和复制，因此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。

为了降低以上问题对新型智算中心整体运行效率的影响，需要引入基于计算总线协议的统一内存池（如 CXL，Compute Express Link）。通过构建统一内存池技术，实现一致性的内存语义和空间寻址能力，显著提高 CPU、GPU 与 Cache、DDR、HBM 等缓存、内存及显存系统的整体效率，从而支持更复杂、灵活的计算模型。

为尽早实现内存池化技术应用，以 CXL 协议为例，应重点在以下几个方面进行增强：

第一，尽快完善满足内存池化技术的计算总线协议及子协议实现。完整、高效地实现 CXL.io 和 CXL.mem 协议，为设备之间的 I/O 通信和内存访问提供通道，优化数据传输和复制机制，降低内存池化引入的额外性能损失，确保系统高效运行。

第二，加快 GPU、AI 加速卡基于 CXL 或计算总线协议实现内存一致性机制。引入内存池技术将减少数据在计算和存储设备之间协议转换频度，通过实现内存一致性机制，优化内存、显存、缓存之间的一致性算法，确保共享内存中的数据同步更新，使得设备之间数据具有一致性和可用性。同时，实现健壮的纠错机制，确保内存池系统稳定可靠运行。

第三，加快制定多异构设备与内存池之间的统一接口，并具备隔离保护能力。提供多异构设备之间的协同工作接口，聚焦设备间高效协作和共享计算能力，减少数据传输和复制所带来的延迟和能耗。同时，强化安全措施，确保只有授权的处理器能访问内存池，防止访问冲突。

因此，推动产业基于计算总线协议构建的新型智算中心内存池技术，将在智算发展的道路上掀起一场重要的变革。内存池技术的应用将使得 CPU、GPU/AI 加速卡等异构设备共享统一内存，简化数据传输和管理，显著降低系统的复杂性和能耗。内存池技术的发展将为 AI 领域带来更高效的创新机会。

6 新平台——融通无限生态

智算平台的关键在于对智能算力进行高质量管理，使能资源效率更高、计算性能更优、业务入驻更易、算力协同更广。智算平台高质量管理有四大关键技术可供选择，从近期来看，用于优化资源效率的池化技术以及提高计算性能的分布式训练框架技术较为成熟，应在新型智算中心建设中引入并持续演进；从中远期分析，应在进一步培育、完善国内自主分布式训练产品的同时，加快推动算力原生技术成熟以降低业务准入门槛，深入研究跨节点分布式训练技术实现离散异构智算资源整合。

6.1 池化技术优化资源使用效率

传统智算中心的 GPU 利用率面临巨大挑战，据公开数据显示，已有智算中心，GPU 平均利用率不超过 30%，比如 AWS re:Invent 2018 公布数据显示其平均 GPU 利用率为 20%、Facebook 2021 年机器学习负载分析报告显示 GPU 利用率不足 30%、英伟达 GTC2022 公

布数据显示 Google 云的 GPU 利用率为 25%。GPU 利用率较低通常是由于其资源分配方式导致，传统智算中心的 GPU 资源分配以整卡分配或虚拟化分配为主，粒度较粗，资源静态绑定，且多规格业务共存导致资源碎片化。

智算资源池化平台以“软件定义”的方式，提供四大资源敏捷管理功能，优化资源效率降低总体购置成本。

●**化整为零：**改变传统的整卡分配、一虚多的粗放式分配方式，使能精细化分配能力，根据 AI 任务需求做到 1% 的细粒度资源按需供给；

●**隔空取物：**基于高速无损网络，跨节点调取智算资源，使 CPU 及 GPU 高度解耦，进一步降低碎片化率；

●**化零为整：**整合分布在多机上的零散资源，汇聚成为大模型业务可使用的资源，使资源可高效分配；

●**变静为动：**改变传统的资源静态绑定的机制，使能资源可以根据负载变化动态分配、回收，多任务间可以峰谷互补，全局资源可以适度超分，促进资源效率提升。

当前业界已经具备成熟的智算资源池化产品，以趋动科技的 OrionX、VMware 的 BitFusion 等产品为典型代表，该类产品基于主流的 Kubernetes 容器管理技术构建池化管理能力，通过云原生技术与 GPU 池化技术的“强强联合”，为资源敏捷管理提供有效支撑。

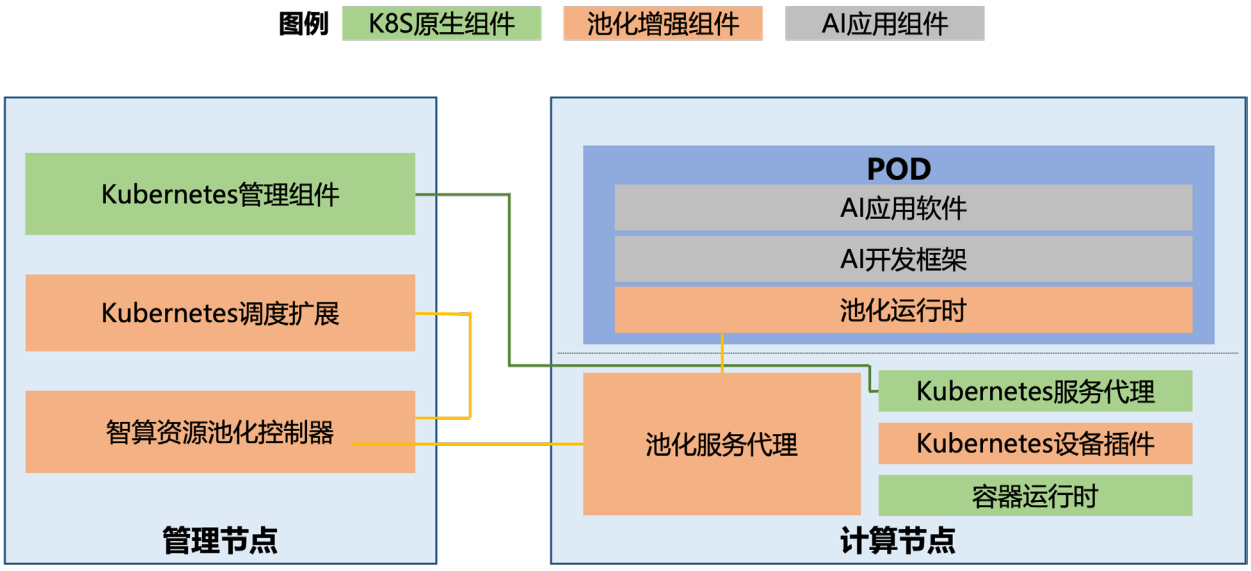


图 3-17 智算资源池化平台

如 3-17 图所示，Kubernetes 作为池化平台的技术底座，主要承担 CPU 的管理调度和作为 AI 任务载体的容器 POD 的生命周期管理功能，通过对 Kubernetes 调度能力的扩展，将 GPU 等智算的管理功能转移至独立的池化控制器执行。而在业务侧，在容器 POD 内植入完全仿真 GPU 卡的原生运行时的池化运行时组件，AI 应用可以像在真实环境中一样运行无感知，通过池化运行时劫持 AI 应用对 GPU 的访问 API 并转交池化服务代理执行，再由池化代理配合池化控制器实现四大敏捷管理功能。

上述智算资源池化技术有效优化了 GPU 等智能算力资源的管理效率，然而 GPU/AI 芯片种类繁多，其原生运行时又相对活跃、升级频繁，对原生运行时进行仿真以构建池化运行时的工作较为复杂，开发量、维护难度较大，不利于技术演进和运营运维。

另外一种流派的池化技术可规避基于 API 劫持技术所面临的问题，该类技术将 API 劫持转移至更底层的驱动层面实现，该位置涉及的接口更少，可大幅度降低仿真工作复杂度，以 VMware 的 Radium、阿里云的 cGPU、腾讯云的 pGPU 等产品为典型代表。

这是一种完全与 GPU 无关的设备虚拟化和远程处理方法，允许在没有显式软件支持的情况下启用新的硬件体系结构。该项技术分为前端和后端，前端监视应用程序对驱动的调用，拦截至后端处理，后端则按应用程序申请的数量分配资源，或将应用程序拆分到多台机器上运行，在保持代码、数据和执行环境一致性的前提下使用这些机器上的智算资源，从而实现资源的敏捷化管理。与 API 劫持技术直接介入到 AI 应用访问资源的流程、需要仿真原生运行时的 API 接口的方式不同，应用程序监视器不介入到 AI 应用访问资源的流程，而是通过更底层的系统调用隐含而广泛的支持更多种类、型号的硬件和新的运行时功能，其实现方式与特定的运行时 API（如 CUDA）无关，具备更加强大的通用性和兼容性。

上述池化技术实现较为复杂，但灵活性较高，然而 GPU/AI 加速卡驱动的接口多为不透明，对驱动调用的劫持面临一定的兼容性问题，且存在一定的法律风险。

两种方案在集成实现难度、性能表现、升级适配等方面各有优劣，用户需根据实际应用情况选择。芯片多样化、生态不融通而导致的竖井问题是制约技术发展的根源，在此呼吁产、学、研各界合作伙伴精诚合作，凝聚共识，在此“一芯难求”的时代，聚焦可用算力、高效利用，共促智算产业有序发展。

6.2 算力原生融通多样算力生态

新型智算中心需要集结泛在、多样的计算系统，形成一体化的灵活服务能力，随时随需为智能应用适配计算资源，以满足人工智能等产业迅猛发展所需的巨大算力需要。

当前，多样异构计算系统一体协同运用面临严峻的竖井化生态挑战。各厂商围绕自身硬件特性构建相对独立且排他的工具链系统，适配集成各类 AI 框架形成分支版本，构成“中间件 / 框架 + 工具链 + 硬件”紧密协同的长链条式智算生态；各厂商间互不兼容，致使上层智算应用与特定系统的锁定，难以在多个竖井生态系统间迁移部署，使算力运营商所集结的多样算力无法为智算应用呈现出一体化的资源，制约算力资源的高效运用，亟需融通业界生态竖井，屏蔽底层算力资源复杂性，使能应用无感迁移部署。

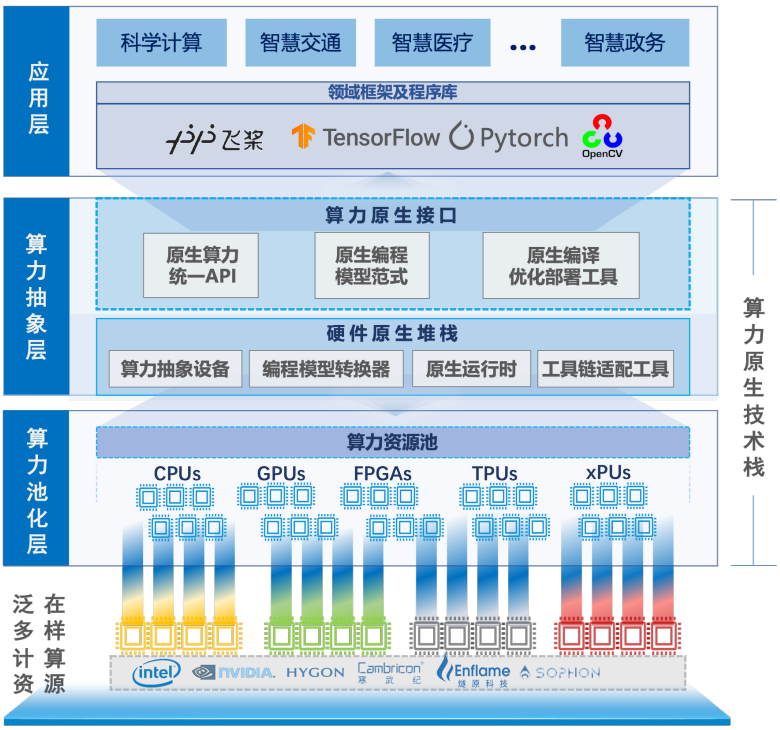


图 3-18 “芯合”算力原生跨架构平台

算力原生技术应运而生。一是为动态变化、多样泛在的异构计算架构构建统一的算力设备抽象模型；二是为开发者建立统一的编程模型及范式，实现转译机制及性能综合优化；三是为各类算力厂商软硬件栈建立动态适配、统一纳管及任务映射机制。从而以新型技术机制解耦各厂商紧锁定式工具链运行模式，从根本上解决上层应用对单一厂商工具链的依赖问题，屏蔽混合算力环境下多类底层硬件差异，实现应用的动态自适应跨架构迁移部署运行。实现“同

一应用一次封装、屏蔽差异跨芯迁移、智算应用一体部署”的能力目标。

以算力原生技术为依托，中国移动联合产业伙伴全速打造可部署于混合异构集群内的“芯合”算力原生跨架构平台 [14]，构建跨架构编译器、原生运行时、集成开发环境等关键产品组件，为开发者用户提供云端可随时接入的统一入口，实现异构算力资源监控、跨架构程序开发、原生程序管理、调试部署等工作的全生命周期解决方案。当前正加速推动组件研发、系统适配、联合测试、标准建设等工作，希望与产、学、研各界合作伙伴精诚合作，凝聚共识，共同推进算力原生技术成熟，繁荣智算产业全“芯”生态。

6.3 分布式训练提升模型训练效率

分布式训练框架是搭建在大型算力资源池基础上的用于并行处理深度学习模型分布式训练任务的一组工具集合，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、可靠和快速的深度学习模型训练，提高模型的准确性和效率。

当前业界分布式训练框架呈现百花齐放现状，一方面是以硬件厂家主导研发的框架，这些框架特点在于与其硬件配合度较高，能够较好地发挥硬件本身优势，包括英伟达 Megatron、华为 Mindspore 等；另一方面是传统主流 AI 框架研发的分布式训练库，包括微软 DeepSpeed 等；同时比照国外业界主流的框架，国内各厂商也研发了对应的框架，包括百度 PaddlePaddle 等。这些框架均针对各自底层硬件需求、并行策略接口设计、算法优化等设计了特定功能。

综合分析当前主流分布式训练框架及技术，该领域仍面临部分挑战和难点，主要包括：

●**通信开销：**在分布式训练中，各个计算设备之间需要频繁地进行通信，以传输模型参数和梯度信息。这样的通信开销如何与计算匹配设计是模型训练调优的关键。

●**同步问题：**分布式训练需要对不同设备上的模型参数进行同步，以保持全局模型的一致性。然而，设备之间的计算速度可能不同，导致一些设备比其他设备更快地完成计算，从而产生同步问题，需要设计一些同步策略，如梯度累积、异步更新来解决这个问题。

●**容错性：**在分布式训练中，经常发生因设备故障或通信错误影响训练效率。需要采取合理的机制处理设备故障，如断点续训、备份数据等。

●**调试和监控：**分布式训练的调试和监控对于训练出现问题时的追踪和定位非常关键，这对于大规模的分布式系统来说是一个挑战。

尽管当前存在以上技术挑战，但研究人员和工程师们仍在不断改进和优化分布式训练技术，使其能够适应更大规模和更复杂的深度学习模型和数据集，通过在硬件、算法、网络和系统方面持续创新，进一步推动分布式训练技术的发展。未来分布式训练框架演进的总体趋势包括：

●**自动化、简洁化。**框架开发者通过提供更加简单易用的接口、工具，逐步使能半自动化、自动化分布式并行训练，提高算法研发效率，助力用户高效编写分布式并行代码，并简化分布式训练配置、调试等过程；

●**弹性和动态性。**未来的分布式训练框架将能够根据实际需求进行资源的动态分配和释放，实现在云计算环境下更加高效地利用计算资源，满足不同规模和需求的训练任务；

●**跨平台训练任务迁移。**为了满足来自用户的多样化需求，分布式训练框架将逐步支持跨平台的训练能力，用户可以在不同的硬件设备间无缝切换和迁移训练任务；

●**支持超大规模模型训练。**随着数据量和模型规模的不断增加，分布式训练框架将进一步优化设计并行训练策略以满足超大规模模型训练需求，并考虑支持跨集群分布式训练解决单集群算力发展规模存在上限的问题，其可能涉及底层通信操作、资源利用策略等优化机制。

●**边缘侧训推。**随着边缘计算等技术的发展，未来分布式训练框架将面向边缘侧设备计算、存储资源限制等瓶颈，通过引入模型压缩、轻量化、异步训练等技术，满足边缘计算场景下的大模型实时训推需求。

●**更好的容错能力和鲁棒性。**分布式训练框架将不断改进其容错和鲁棒能力，以应对计算故障或通信中断等问题，包括设计更优的任务检查点机制、容错调度策略等技术，确保训练过程的稳定性和可靠性；

●**多模态、多任务训练能力。**随着多模态数据和多任务学习的兴起，分布式训练框架将逐步支持同时处理多个数据模态和执行多个任务，通过提供可适应多模态数据输入和输出的接口和算法，并探索多任务学习的优化策略，提高模型泛化能力和效果；

●**在线学习和增量学习。**在某些应用场景中，数据可能是不断产生和变化的，需要框架支持

在线学习和增量学习的能力。未来的分布式训练框架将支持在线学习的模式，能够在训练过程中动态地接受新数据并实时进行模型更新。

整体来看，未来的分布式训练框架将聚焦于自动化、弹性、跨平台支持、大规模模型训练、跨集群训练、边缘训推、容错可靠性等方面不断优化完善，简化用户在开发大模型过程中的资源开销，通过功能完善、算法改进等方式提供更加高效、易用的模型训练工具。

6.4 跨域分布式调度促进广域资源利用

当前，智算资源呈现异构和离散状态，各地智算中心部署的 GPU/AI 加速卡、网络配置等均不相同。此外，受限于机房空间和供电，部分智算中心资源不足无法满足大模型训练需求，部分智算中心则存在大量闲置资源没有充分利用的情况。如何有效利用海量而分散的数据和算力，实现高性能、高可靠的跨域分布式并行训练，将成为推动多智算中心算力协同、促进算力共享与协作的关键。

跨智算中心的分布式并行训练目前已成为学界一大研究方向，旨在通过对多智算中心资源的统一纳管和调度，实现跨域的模型训练，随着智算中心内部算力和带宽的飞速提升，跨域分布式训练的通信效率已成为主要性能瓶颈：

1）跨智算中心可用传输带宽有限，周期性的大流量通信引发通信瓶颈。对于多智算中心间的网络互联，位于同一城市的可采用密集型复用技术和裸光纤直连实现物理链路的互联互通，但异地场景传输距离远不具备光纤直连条件，通常使用广域网专线连接。跨域大模型训练需要周期性同步模型参数，随着模型规模增大，智算中心间通信将产生严重的网络拥塞和性能瓶颈。

2）跨智算中心算力和网络资源差异分布且动态变化，易产生同步阻塞，拉低系统效率。不同智算中心算力和网络配置不同，将造成计算和传输步调不协调。对于跨域训练问题，算力低、带宽小的计算节点将拖慢模型同步的完成时间并拉低系统训练效率，产生木桶效应。此外，广域网的带宽也要分配给其他通信业务，带宽资源实时竞争，资源的动态时变特性使模型同步更加难以估计，进一步加重同步阻塞。

综上，优化模型传输带宽开销和确保模型训练性能是跨域模型训练需重点解决的问题，需要

在系统架构、传输协议和顶层算法等方面进行联合优化设计：

●**系统架构层面，**研发跨域分布式训练框架和全局资源调度平台。统一训练框架可形成统一的资源抽象，以支持所有的并行策略（数据并行、流水并行、张量并行）及策略混合，同时具备高效、通用和硬件感知等能力，实现训练任务拆分、任务下发和参数合并等。全局资源调度系统可实时获悉不同智算中心的硬件资源情况，监控节点间带宽资源变化，并使用断点续训等手段缓解硬件故障和带宽阻塞的影响。

●**传输协议方面，**设计面向模型训练的梯度差异性的传输协议，减少跨域通信带宽需求。当前的分布式机器学习系统普遍采用精确传输服务协议，无差别地将所有梯度从发送端完整地传输到接收端（如 RoCE 协议），并未考虑梯度的差异化传输需求。然而，不同梯度对模型收敛影响不同，无需传输所有梯度信息即可完成模型训练。因此，设计一种根据梯度贡献的传输协议可提升跨域场景下的高效通信。

●**算法设计方面，**探究模型的压缩传输技术，减少域间模型数据量的传输需求，缓解通信瓶颈。模型压缩可以显著减少每次模型同步传输的数据量，典型的方案包括量化和稀疏，量化是用有效的数据表示压缩参数或梯度，减少每个值的占用位数，稀疏则是通过筛选部分关键梯度参与梯度聚合与模型更新，避免发送不必要的信息。设计满足跨域带宽需求的模型压缩传输算法，可有效减少域间带宽需求，缩短模型训练时间。

7

新节能——实现可持续发展

高性能智算服务器功率密度正在急剧上升，CPU 从 150W 增加到 300W 以上，GPU 则增加到 700W 以上，一台配置了 8 卡 A800 的智算服务器功耗约为 6.5KW，是通用服务器的 10 倍以上，这意味着服务器散热量会大大增加，单机柜的功率密度和散热量也大幅增长。与此同时，半导体元器件的温度每升高 10 度，可靠性就降低一半。据统计，电子器件失效中 55%由温度过高或温度不均造成，芯片的极限温度为 85 度，超出这个阈值，芯片就会降低性能和功耗来保证其稳定可靠的运行。

液冷技术可以有效地将热量从设备中导出，相比空气能够传热更快（提升 20~25 倍），能够带走更多热量（提升 2000~3000 倍），有效降低设备温度。采用液冷技术一方面可以提高数据中心的设备部署密度，实现空间资源的高效利用（液冷机柜密度是传统风冷机柜的约 3~4 倍，相同算力下节省机房面积约 75%）；另一方面将有助于提高芯片可靠性，保证芯片在最大电压和频率下连续运行，提升芯片性能。

液冷方案包括冷板式、浸没式和喷淋式三种技术。冷板式液冷是一种非接触式液冷方式，液体无需接触发热的器件，通过装有液体的冷板导热，借助液体循环带走热量。浸没式液冷为接触式冷却，将发热器件完全浸没在冷却液中，发热器件与液体直接接触并进行热交换。根据工质是否产生相变又分为单相液冷和相变液冷。喷淋式液冷是将冷却液直接喷淋到发热器件表面或与发热器件接触的扩展表面上，吸热后排走，再与外部冷源进行热交换。冷板式和单相浸没式是目前主流方案。冷板式液冷部件兼容性强，机房改造、运维工具和运维习惯上相对于风冷改动较小，可实现平滑过渡。单相浸没式具有更低的 PUE，但存在冷却液成本高、与现有基础设施不兼容、生态不完善等问题，随着国产冷却液性能的提升，产业生态不断成熟，浸没式液冷应用场景将进一步拓展。**总体来看，冷板式液冷和单相浸没式液冷各有优劣，短期内将以冷板式液冷为主，长期来看冷板式和单相浸没式液冷将共存发展。**

考虑到大模型训练场景散热和可靠性需求，兼顾机房空间、设备部署、成本以及产业成熟度等因素，大规模引入冷板式液冷技术可能面临如下挑战：

1）统一标准问题：液冷系统涉及到的部件之间兼容性存在标准缺失，各家服务器设备、冷却液、制冷管路、供配电等产品形态各异、接口不同，导致产业竞争不足，采购成本高，影响产业高质量发展。

2）可靠性问题：除了服务器本身，冷却液流经的管路也存在腐蚀和泄露的风险，冷却液对管路的腐蚀会引起管道的损坏，进而导致泄露，因此防腐蚀是液冷系统设计的关键。此外，泄露后的故障隔离措施和低温环境下液体的冻结风险也需要重点关注。

3）液冷系统的监控和运维：相比风冷系统，需重点关注新增 CDU 设备和液冷管路的维护，以及冷却液的监测和维护。

针对上述问题，有如下解决方案供行业参考：：

1）建议采用液冷整机柜和 CDU 解耦方式，将液冷整机柜接口，包括液体压差、流速、温度、管路接口型号等参数，形成行业标准，不同厂商的液冷机柜和同型号 CDU 对接，实现异厂家液冷整机柜服务器共机房部署，降低数据中心建设、运营成本。

2）在冷却液中添加含有防腐蚀、防冻功能的缓蚀剂，比如乙二醇溶液等，降低管道的泄露以及液体冻结风险。增加漏液导流结构设计，一旦泄露后具有故障隔离功能。

3）采用集中式 CDU 部署，减少运维工作量，故障维护不停机，机房空间利用率高。

面向中远期，我们将继续推动液冷技术发展和落地应用：

1）推动服务器和机柜解耦：通用冷板式液冷服务器已被行业广泛接受，生态趋于完善，例如机柜，冷板，快接头，manifold，CDU，漏液监控和处置措施等，各种团体标准都在制定中。基于快接头标准，双路通用计算服务器可以实现机柜和服务器之间的解耦。但是在 NICC 场景下，产业生态标准化不足，产品设计差异较大，冷板能力、流阻、流量差异更大，机柜和服务器解耦难度较高，导致产品竞争不充分，采购成本高，同时给设备运维带来困难。我们将推动制定液冷服务器和机柜解耦相关标准，促进液冷产业全链条发展。

2）运维和管理优化：需要对液冷新增的 CDU、液泵、液冷的 ICT 设备进行统一管理；运维方面，必须采取健康防护措施，对冷却液的更换和排放必须遵循化学品的安全技术要求，运维人员的技能水平也必须紧跟发展步伐。

3）持续优化 PUE：扩大冷板散热的冷却范围，降低风冷散热的比例。结合高效风冷散热系统，进一步降低冷板式液冷数据中心的散热。协同制冷和散热，合理设计服务器冷却液进口温度、科学分配散热温差、实现服务器及其散热系统的节能最大化。

8 总结和倡议

大模型技术的日新月异对智算底座的升级提出了高要求，由于硬件的迭代周期和成本都要远大于上层软件和算法，统筹考虑、超前布局基础设施技术方案尤为重要。本白皮书从新互联、新算效、新存储、新平台和新节能方面提出新型智算中心的技术演进建议，希望与产业链各方达成共识，共同推动智算关键技术成熟和生态繁荣发展，我们建议：

面向新互联，面向百卡级别的高速互联需求，产业应联合打造统一的计算总线协议，实现缓存一致性的数据访问，并提升流量控制、拥塞控制、网络无损、重传等通信和数据传输能力，收敛技术路线，推动国内高速互联技术生态成熟；集群间基于 GSE 打造无阻塞、高带宽、低时延、自动化的新型智算中心网络，向更细粒度的负载分担、端网结合的拥塞控制和基于全局的智能运维三个方向不断演进。

面向新算效，从存算一体化、稀疏化、AI 算子硬件支持、更低推理时延等方面进一步优化

AI 芯片设计；要大力推动存算一体与大模型技术的结合，从算法结构设计、精度需求和先进封装技术三个角度，加速多核、多芯片存算一体架构成熟；按需引入 DPU，加快推进云平台软件、DPU 卡硬件、服务器硬件的标准化，最大化助力算效提升。

面向新存储，结合新型智算中心多元异构的数据特征，打破传统各协议分立的存储架构，共同推进多协议原生融合存储的产品研发、技术成熟和商用部署，积极探索基于 CXL 的统一内存池和跨地域全局统一存储技术方案。

面向新平台，在引入智算池化技术和分布式训练框架的基础上，加快推动算力原生技术成熟以降低业务准入门槛，同步探索跨域分布式训练技术，实现离散异构智算资源整合。

面向新节能，坚定推动液冷技术成熟，解决新型智算中心的散热压力和节能挑战。聚焦液冷服务器和液冷机柜的接口标准，优化液冷环境下运维和管理能力，促进产业链上下游生态成熟和能效利用水平不断提升。

缩略语列表

缩略语	英文全称	中文解释
AI	Artificial Intelligence	人工智能
AIaaS	AI as a Service	AI 即服务
AIGC	AI Generated Content	生成式人工智能
API	Application Programming Interface	应用程序编程接口
CC	Congestion Control	拥塞控制
CDU	Coolant Distribution Unit	冷量分配单元
CIM	Computing In Memory	存内计算
CMOS	Complementary Metal Oxide Semiconductor	互补金属氧化物半导体
CNN	Convolutional Neural Networks	卷积神经网络
CPO	Co-packaged Optics	光电共封装
CV	Computer Vision	计算机视觉
CXL	Compute Express Link	计算总线协议
CUDA	Compute Unified Device Architecture	英伟达通用并行计算架构
DCQCN	Data Center Quantized Congestion Notification	数据中心量化拥塞通知
DDR	Double Data Rate SDRAM	双倍速率的 SDRAM
DGSQ	Dynamic Global Scheduling Queue	动态全局调度队列
DP	Data Parallel	数据并行
DPU	Data Processing Unit	数据处理单元
DSP	Digital Signal Processing	数字信号处理
EC	Erasure Code	纠编码
ECMP	Equal Cost Multi Path	等价多路径
ECN	Explicit Congestion Notification	显式拥塞通知
FC	Fully connected	全互联拓扑
FEC	Forward Error Correction	前向纠错
GSE	Global Scheduled Ethernet	全调度以太网
GSF	Global Scheduled Fabric	全调度交换网络
GSOS	Global Scheduled Operating System	全调度操作系统
GSP	Global Scheduled Processor	全调度网络处理节点
HBM	High Bandwidth Memory	高带宽内存
HoL	Head-of-Line Blocking	队头阻塞

缩略语	英文全称	中文解释
IaaS	Infrastructure as a Service	基础设施即服务
IB	InfiniBand	“无限带宽”技术
LPO	Linear-drive Plugable Optics	线性驱动可插拔光模块
MaaS	Model as a Service	模型即服务
MoE	Mixture of Experts	专家并行
MP	Model Parallel	模型并行
MRAM	Magnetoresistive Random Access Memory	非易失性的磁性随机存储器
NICC	New Intelligent Computing Center	新型智算中心
NOS	Node OS	节点操作系统
NPU	Neural network Processing Unit	神经网络处理器
NVMe-OF	NVMe over Fabrics	基于架构的非易失性内存标准
OAI	Open Accelerator Infrastructure	开放加速器基础设施
OAM	OCP Accelerator Module	OCP 加速模组
OCP	Open Compute Project	开放计算项目
PCIe	Peripheral Component Interconnect express	高速串行计算机扩展总线标准
PFC	Priority-based Flow Control	基于优先级的流量控制
PIM	Processing In Memory	存内处理
PKTC	Packet Container	报文容器
PNM	Processing Near Memory	近存处理
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
RoCE	RDMA over Converged Ethernet	融合以太网承载 RDMA
RDMA	Remote Direct Memory Access	远程直接数据存取
RRAM	Resistive Random Access Memory	可变电阻式存储器
SDN	Software Defined Network	软件定义网络
SRAM	Static Random-Access Memory	静态随机存取存储器
UBB	Universal Baseboard	通用基板

参考文献

[1]Maintaining American Leadership in Artificial Intelligence, https://www.federalregister.gov/documents/2019/02/14/2019-02544/maintaining-american-leadership-in-artificial-intelligence

[2]S.1260 - United States Innovation and Competition Act of 2021,

https://www.congress.gov/bill/117th-congress/senate-bill/1260

[3]2030 Digital Compass: the European way for the Digital Decade. https://eufordigital.eu/wp-content/uploads/2021/03/2030-Digital-Compass-the-European-way-for-the-Digital-Decade.pdf

[4]Summit.

https://www.olcf.ornl.gov/olcf-resources/compute-systems/summit/

[5]Polaris, Argonne Leadership Computing Facility

https://www.alcf.anl.gov/polaris

[6] Mai G, Huang W, Sun J, et al. On the opportunities and challenges of foundation models for geospatial artificial intelligence[J]. arXiv preprint arXiv:2304.06798, 2023.

[7] Huang Y, Cheng Y, Bapna A, et al. GPipe: Easy scaling with micro-batch pipeline parallelism[J]. arXiv preprint arXiv:1811.06965, 2018.

[8] Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. Megatron-lm: Training multi-billion parameter language models using model parallelism. arXiv preprint arXiv:1909.08053, 2019

[9]OCP 开放加速器基础设施项目

https://www.opencompute.org/wiki/Server/OAI

[10] Kaplan J, McCandlish S, Henighan T, et al. Scaling laws for neural language models[J].

arXiv preprint arXiv:2001.08361, 2020.

[11] 全调度以太网技术架构白皮书，中国移动研究院

[12] 存算一体白皮书，中国移动研究院

[13]ZHU Haozhe, JIAO Bo, ZHANG Jinshan, et al. COMB-MCM: Computing-on-memory-boundary NN processor with bipolar bitwise sparsity optimization for scalable multi-chiplet-module edge machine learning[C]. 2022 IEEE International Solid-State Circuits Conference (ISSCC), San Francisco, USA, 2022: 1 – 3.

[14] 面向智算的算力原生白皮书，中国移动研究院