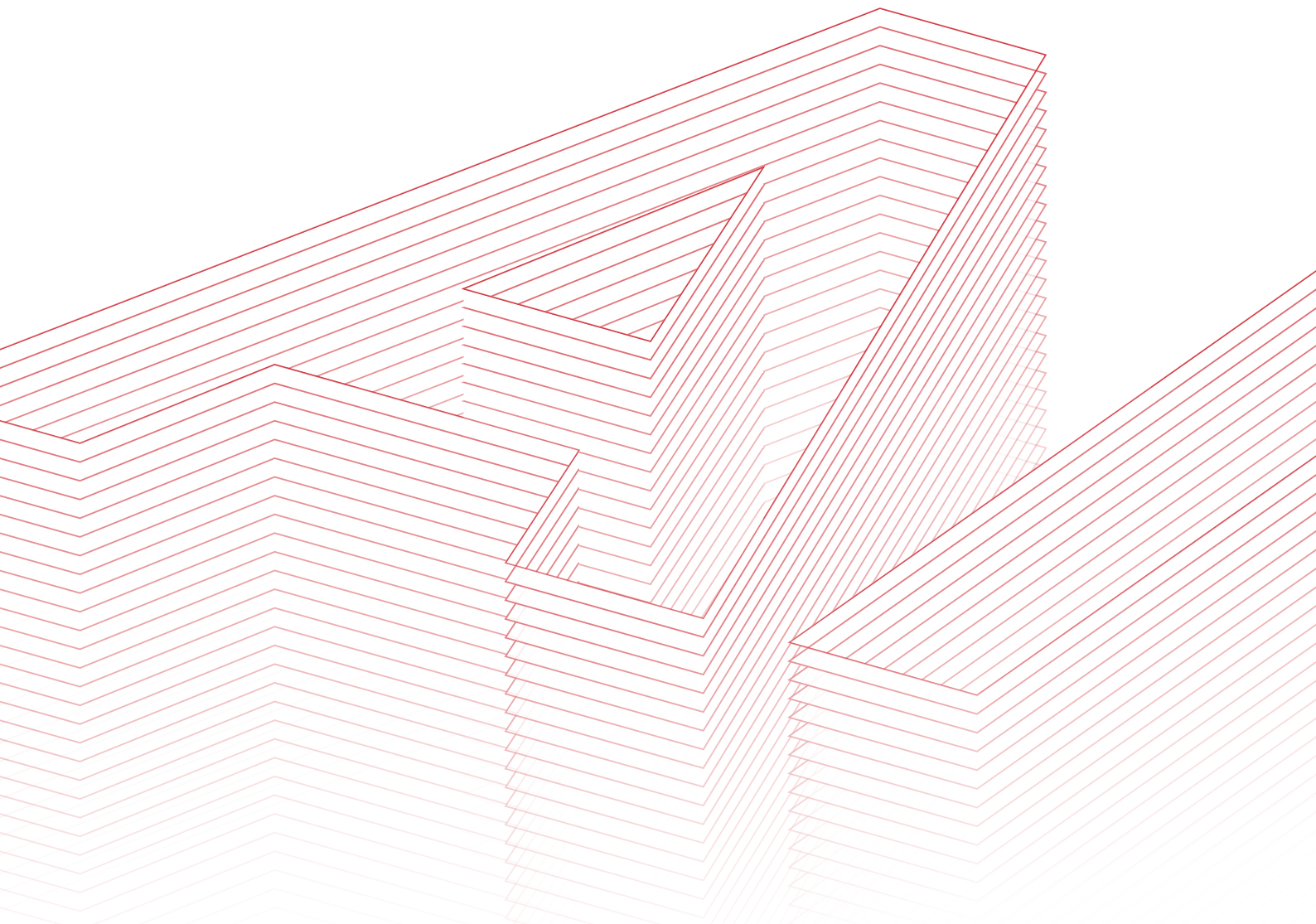
AI-Native
技术与实践白皮书

创原会 CLOUD NATIVE ELITE CLUB | HUAWEI



联合出版单位

创原会、华为云

白皮书编撰组

主编

顾炯炯 华为 Fellow、华为云首席架构师

特邀作者

朱熠锷 金山办公 助理总裁

王云峰 值得买科技 CTO

代 迪 美宜佳 CIO

吴鸿钦 美宜佳 技术总监

刘福东 汉得信息 董事会秘书&副总裁（平台产品）

编撰成员

叶 涛、曹伟朋、王勇桥、徐传飞、陈 光、郜忠华、付 萌、朱 磊、伍华涛、宋江娴、马会彬、蒋 昊、曾 凯、张秦涛、蒋东生、受春柏、陈 衍、唐盛军、练韵文、黄哲思、王健楠、李 昆、王 晨、许田立、邓红斌、李向阳、徐礼锋、廉 莲、陈懿斌、徐云昆、党 倩、石苏龙、林歆远、常建龙、李智华、刘建锋、单一舟、康永红、孙彬彬、文永新、刘 博、黄卫立、毛 杰、 罗 斌

特邀顾问

张宇昕 创原会荣誉理事长

董理斌 华为云Marketing部部长，创原会副理事长

发布时间：2025年12月

序

当今世界，人工智能（AI）正以前所未有的速度重塑人类社会的生产方式和创新范式。在这场深刻的第四次工业革命浪潮中，通用人工智能（AGI）的曙光初现，以OpenAI、Gemini、DeepSeek、Kimi等为代表的生成式大模型持续突破技术边界，加速向人类认知水平迈进。2025年，AI大模型将完成从“智能对话助手”向“全能任务执行者（AI Agent）”的跃迁。上述AI的进化不仅驱动生产力的指数级跃升，更深刻重构着全球竞争格局——率先实现AI技术与产业深度融合者，将在智能时代的经济版图中占据战略制高点。

这场智能化革命的核心，在于技术范式正经历从“AI+产业”到“产业×AI”的本质性跨越。传统的AI应用如同工业时代的蒸汽机，仅在特定环节释放局部动能；而AI-Native架构则如同电力革命，其核心价值在于重塑企业的基础设施基因、创新模式与价值网络，驱动企业乃至整个产业生态向真正意义上的“智能有机体”演进。

华为云通过昇腾云服务与CloudMatrix算力平台的软硬协同创新，构建了端到端的国产化智能算力体系。其超节点架构突破性地实现计算与存储资源的超高速对等互联，将大模型训练性能提升68%、推理效率提升30%，为AI-Native架构提供澎湃动力。这一突破不仅解决了传统算力架构在智能密度指数级增长下的性能瓶颈，更通过“随取随用”的云服务模式，让企业能够灵活应对大模型训练、实时决策等高负载场景。昇腾超节点的诞生，标志着中国在智能算力领域构建了自主可控的第二平面，为关键行业提供了安全可靠的智能基础设施底座。Agentic AI的群体智能革命是华为云AI-Native架构的另一核心突破。其多Agent协同框架通过角色定义引擎、任务分解器和动态编排中枢三大模块，构建了可进化的智能生态系统。每个Agent既是特定领域的“技能专家”，又能通过联邦学习形成知识共享网络。这种“超个体智能”架构突破了单体Agent的极限，使企业能够通过多Agent的自主协作，完成复杂系统的智能决策与动态优化。

为系统化梳理AI-Native的技术体系与实践经验，华为云联合创原会的行业领袖、技术专家与学术机构，历时一年深入研讨，最终凝练成此部白皮书。本书不仅是一份技术指南，更是一份面向未来的宣言。我们期待，本白皮书能为各行各业的智能化转型提供可落地的参考框架，助力企业将AI的应用从“降本增效的工具”，升级为“重构新质生产力的引擎”，从而开辟企业创新与增长的新曲线。

智能时代的浪潮奔腾不息，AI-Native的征程才刚刚启航。我们始终坚信，技术的价值在于普惠——我们愿以自身在云计算、AI、行业数字化领域的深厚积累，与全球客户、开发者及生态伙伴携手，共同探索AI-Native的无限可能。让我们以开放的心态拥抱变革，以创新的勇气定义未来，共同迈向“万物皆智能、千行皆重塑”的新纪元！

目 录

01

前言

1.1 背景

回顾2025年，AI领域的发展可谓“风起云涌，高潮迭起”，从年初的DeepSeek V3/R1开源大模型异军突起一举打破硅谷大模型巨头的垄断，再到DeepSeek-OCR对超长上下文的颠覆式创新，GPT/Claude/Grok/Gemini竞相发布新品，发布不断刷新了大语言模型性价比和推理能力的上限，而大模型的应用也从聊天对话和内容生成全面升级为目标驱动可独立思考规划并调用工具完成复杂任务智能体，正式开启了“Agent元年”，企业开始扎堆投入Agentic应用智能化改造，而多模态大模型及世界模型在自动驾驶、机器人具身智能以及媒体娱乐行业的应用落地也不断取得新的突破。

由此可见，生成式人工智能正在以革命性姿态引领第四次工业革命，其作为AI原生系统的核心驱动力，正在重塑全球产业格局的底层逻辑。从感知智能到认知智能的跨越式演进，不仅使AI系统具备了类人的环境理解与自主决策能力，更推动技术范式从预定义规则向“数据驱动-AI使能-算力支撑”三位一体的根本转变。这一进程催生了以“智能内生”为本质特征的AI原生系统——它们不再依赖人工规则配置，而是通过持续学习形成动态优化能力，彻底颠覆了传统IT架构的设计范式，标志着AI原生时代的正式来临。

在数字化与数智化的双重演进中，生成式AI已从技术工具升维为生态级核心引擎，其动态适应能力正在重构企业系统的认知架构。AI原生系统展现出从“执行指令”到“理解意图”的质变能力，这种基于生成式AI的认知跃迁，使得业务流程能够实现自主优化与自我迭代。这种转变倒逼基础设施向“AI原生就绪”形态进化：算力需支持大模型分布式部署以及大/小模型/Agent协同，数据架构必须适应实时处理与反馈闭环，开发模式转向以数据和AI驱动为核心的新范式。这些变革共同构成了AI原生时代的基础设施标准，为通用人工智能（AGI）的演进铺设了技术通路。

当生成式AI的技术成熟度跨越临界点，AI原生系统将成为所有数字化建设的默认选项。这种转变不仅是技术架构的更替，更是生产关系的革命——从生产流程的百倍效能跃升，到商业模式的全局重构，AI原生思维正在重新定义价值创造的方式。在AI原生时代，企业竞争力将取决于其系统“智能内生”的深度：能否实现需求自感知、策略自生成、效果自优化的完整智能闭环。这场由生成式AI驱动的认知革命，终将推动人类社会从信息化、数字化迈向真正的智能化文明，为超级智能（ASI）时代的到来奠定范式基础。

1.2 白皮书的目的

1) 阐明AI原生技术的内涵与价值

本白皮书旨在构建AI-Native技术的认知坐标系：从技术维度解析其“数据-算法-算力”三位一体的架构特征，从商业维度揭示其“感知-决策-执行”闭环创造的价值飞轮，从战略维度阐释其对企业数字化转型的范式重构作用。区别于传统嵌入式AI的“功能补丁”模式，AI-Native系统具备三个核心特征：架构层面的智能内生性（Intelligence-Native）、数据层面的自进化能力（Autonomous Evolution）、业务层面的价值涌现性（Emergent Value）。本白皮书将据此给出AI-Native的定义，介绍典型的AI-Native技术架构，并分享华为云在AI-Native资源层、AI-Native OS层、以及AI-Native应用层的创新工作。

此外，本白皮书还将深入分析AI-Native技术在不同领域和行业中的应用场景，探讨其带来的技术优势与业务价值。从提升企业运营效率、加速产品创新、增强决策能力到推动产业结构升级，AI-Native技术为各行各业带来了前所未有的发展机遇。通过对AI-Native技术的定义、特征及其应用场景的详细解析，本文旨在帮助企业、学术界和政策制定者理解AI-Native的潜力及其长远影响。

2) AI原生技术参考架构与最佳实践分享

随着AI-Native技术的逐渐普及，许多领先企业已经开始在各自的业务中实施AI-Native架构。AI-Native架构不仅仅是技术的集合，它是通过深度的技术整合，实现AI与企业需求、业务流程和战略目标的高度契合。通过结合当前行业内的最佳实践，本文将分享一些AI-Native架构的设计思路、关键技术以及应用案例。

具体而言，本部分内容将从三个维度展开：首先，在云基础设施层面，将介绍华为云在AI-Native架构中的创新实践，包括多元算力AI超节点、云存储、云网络等核心技术，以及模型使能平台和软硬协同优化方案等；其次，在云服务智能化方面，将分享华为云如何通过AI-Native技术赋能CodeArts、数据库、安全云脑等各类云服务，提升其智能化水平；最后，还将展示AI-Native技术在典型行业中的落地案例，包括金山办公、美宜佳、值得买、汉得信息等企业的成功实践。这些全方位的经验分享，能够帮助企业更深入地理解AI-Native架构的价值，并为其技术实施提供明确的方向指引。

02

AI-Native 技术概述

生成式 AI 的迅猛发展不仅标志着技术能力的跃迁，更催生了一种全新的系统范式——AI 原生。这一理念正在重新定义数字世界的构建方式：当智能不再是被赋予的特性，而是系统与生俱来的核心能力，传统业务架构的价值链将迎来根本性重构。如果说前文揭示了生成式 AI 作为产业变革引擎的宏观图景，那么理解 AI 原生的深层内涵与价值体系，将成为把握这场智能革命关键脉络的认知基石。在接下来的探讨中，将穿透技术表象，剖析 AI-Native 的定义与特征。

2.1 AI-Native的定义与特征

需要明确的是，AI-Native并非一个非黑即白的绝对状态，而是一个标志着应用系统智能水平高低的频谱。一个真正的AI-Native应用，其设计与构建应系统性地体现“AI First”的核心理念，并深度融合数据与知识驱动、自学习、统一模型基座、Agentic行动、以及多元算力支撑等一系列关键特征。这些特征共同构成了衡量AI-Native应用成熟度的标尺，后文将依据这些特征的完整度，定义从L0到L5的六个分级，为评估与实践提供清晰的路径指引。

1) AI First

AI-Native技术的核心理念是“AI First”，即从系统设计伊始便将AI作为核心组件，而非在现有系统中后期集成AI技术。这种设计理念体现了AI在整个系统生命周期中的重要性，旨在将AI能力最大化地融入到系统架构、业务流程、数据流转等各个层面。与传统的“Embedding AI”模式相比，AI-Native技术从架构设计到功能实现都围绕人工智能展开，确保每一个业务环节都能最大限度地发挥AI的优势，提供智能化的解决方案。

各业务领域软件产品及云服务的所有生命周期环节，包括产品与功能规格定义，架构设计，研发过程的开发、测试、发布与运维等各阶段，均需优先思考AI可以做什么，不能做什么，哪些核心功能可以由AI提供，哪些不行：比如基于AI的人机交互，基于AI的核心业务逻辑功能实现，基于AI的需求管理、自动化代码开发测试，应用前后台架构与模型基座之间的组合集成，基于AI的最小化产品选型创新验证与逐步改进等；整体系统的工作流程、核心算法及技术架构，均基于AI驱动的理念与洞察，进行必要的优化甚至重构改造；为确保AI能力在AI Native应用发挥预期的作用与价值，系统中的数据处理、模式识别，以及决策制定等都必须持续和自动化迭代。

AI First意味着应用的“AI内置”，同样也代表了一个体系和机制，贯穿各环节、各角色，如果缺少该“内置AI”能力，系统将不复存在。简单来说以往使用AI能力就是简单调用一些AI能力，如API等，但是内建AI不同，它是一个系统也是一个闭环，真正做到“AI无处不在”。AI First同时也意味着AI在研发流程中的「左移」，即在产品设计、功能和架构设计方面思考AI、使用AI，在一个产品idea涌现初期就使用AI。

2) 数据与知识驱动

AI-Native技术的关键特点之一是高度依赖数据与知识。与传统基于规则的系统不同，AI-Native技术通过对海量数据进行深度学习和模式识别，能够自动从数据中提取有价值的信息，并基于此进行决策与优化。这种数据驱动的方式，不仅能处理传统规则系统难以应对的复杂场景，还能通过不断学习提升决策质量。

AI-Native系统构建了“数据-知识”双轮驱动引擎。数据和知识的融合，使得AI-Native系统能够在面对新情况时进行快速适应。通过持续的数据积累与模型优化，AI-Native技术可以不断增强智能化水平，提升整体业务运营效率。企业通过AI-Native架构，将能够更好地挖掘数据潜力，推动数字化转型，并在激烈的市场竞争中保持优势。

3) 自学习、自适应、自优化

自学习、自适应和自优化是AI-Native技术的重要特点之一。AI系统能够根据实时数据不断进行自我学习，通过模型更新和优化提升决策质量。通过自适应能力，AI可以在不同的应用场景中根据新的数据和反馈调整策略，实现动态响应，而自优化功能则使得系统在长期运行过程中不断提升性能，降低资源消耗，保持较高的运行效率。

这种智能化的特点，意味着AI-Native系统不仅能够在静态环境中完成任务，还能够在复杂、动态的环境中自动适应并优化自身行为。例如，在智能制造场景中，AI-Native系统能够根据生产过程中的实时数据，自动调整生产参数，实现精确控制，从而提高产品质量和生产效率。

4) 以统一基础模型作为智能基座

AI-Native系统的智能化根基在于构建统一的基础模型（Foundation Model），其本质是通过通用性强、泛化能力突出的模型架构，为全场景AI应用提供统一的语义空间和知识表达框架。这种“统一基座”模式突破了传统AI系统中模型碎片化、场景割裂的局限，通过参数共享、知识蒸馏和迁移学习等技术，将通用知识与领域知识深度融合，形成覆盖语言、视觉、决策等多模态的“认知底座”。

统一基础模型在AI-Native系统中的价值体现在三个维度：技术维度，实现了“大模型小场景”的适配，结合模型压缩和参数高效微调，在保持高性能的同时适配边缘侧低算力场景；生态维度，通过模型即插件、指令微调等机制，使开发者能以“乐高式”方式组合基础能力，形成行业解决方案；业务维度，其通过统一语义空间打破数据孤岛，使跨业务线的知识共享效率显著提升。

5) 具备自主性与工具调用能力的Agentic AI

AI-Native系统的先进性，在行为层面体现为具备自主性的智能体（Agent）形态。与被动响应指令的传统AI模块不同，Agentic AI能够理解高层目标，并主动进行任务规划、分解与执行。其核心能力在于自主决策与行动，特别是通过工具调用（Tool Use）和本地知识库查询来扩展其能力边界。

一个AI-Native Agent可以自主检索知识库、调用API、执行代码、操作软件或硬件，从而在复杂的数字与现实环境中完成端到端的任务。这种能力使得应用从“智能助手”升级为“智能执行者”，能够动态适应环境变化，并在多步工作流中实现真正的自动化，极大地提升了系统的自主解决问题能力。

6) 极致性价比的多元算力支撑

AI-Native系统的算力需求呈现出前所未有的复杂性与动态性特征，对算力基础设施提出了革命性要求。算力基座需构建基于超节点架构的多元算力池，兼容CPU/DPU/NPU等异构芯片，并支持千/万卡级并行计算能力，通过对等算力网络实现极致的弹性资源调度。推理场景需依托超节点架构的异构计算加速能力，达成低延时高吞吐的智能服务。这种需求已超越传统云计算资源池化的简单逻辑，需要构建从芯片架构到系统软件的全栈协同优化体系——正如DeepSeek通过算法-编译-硬件的深度协同，实现计算效率的指数级提升。此外，只有当算力基础设施具备对应用特征和算力需求的精准感知以及对多元算力的动态量体裁衣式供给能力的智能特性，才能真正释放AI-Native系统的全部潜能。

在AI原生范式下，算力已从被动资源进化为主动的“智能计算引擎”。这要求基础设施不仅提供基础计算能力，更要能够构建支持万亿参数模型分布式训练的异构计算架构，最终形成“算力-算法-数据”的闭环优化系统，使计算资源能随模型复杂度、业务场景动态调整。这种深度协同优化将算力转化为AI原生的核心生产力，其成熟度直接决定了企业能否在智能时代建立竞争优势——正如大模型性能不仅取决于参数量，更取决于每瓦特算力所能产生的智能效能，而超节点的规模效应与对等算力的协同效率将成为关键决胜点。

综上所述，AI-First的设计哲学、数据与知识的双轮驱动、自学习、自适应、自优化的内生能力、统一基础模型的支撑、具备工具调用功能的Agentic AI自主性、以及极致性价比的多元算力使能，共同勾勒出AI-Native应用的完整画像。然而，必须认识到，在从传统软件向AI-Native演进的过程中，并非所有应用都需要或能够一步到位地具备全部特征。正因如此，引入了L0至L5的成熟度分级模型，其目的正是为了客观地衡量一个应用在AI-Native道路上的所处阶段。在接下来的章节中，将进一步展开介绍AI-Native技术的价值与意义、典型的AI-Native架构设计方法论，以及AI-Native架构成熟度分级体系，帮助读者清晰地定位自身产品，并规划其向更高阶AI-Native形态演进的路线图。

2.2 AI-Native技术的价值与意义

AI-Native技术从业务系统设计之初便将智能能力融入到每一个环节，从而有望在业务运营、功能创新、系统进化、乃至商业模式方面实现全面升级。具体地：

1) 精益化业务运营

在传统的企业运营中，许多业务流程需要依赖人工干预，且容易受到人为因素的影响，导致低效或错误。而在基于AI-Native业务系统中，能够通过数据驱动优化每一个环节，通过自动化流程、优化决策和实时反馈，大幅度提高业务运营效率，实现更加高效、精确的运营。同时，AI系统能够持续监控运营状况，发现潜在问题并进行优化，进一步提升业务运营的效率和质量。AI系统还能够利用数据驱动的洞察力优化业务流程，减少资源浪费，实现精准的需求预测和供应链优化，从而降低成本、提高响应速度，助力企业实现精益化运营。

2) 业务功能创新与增强

AI-Native技术不仅能够优化现有业务流程，还能够为企业带来新的功能创新和业务模式。通过深度挖掘数据价值，AI-Native系统可以发现传统方法无法识别的潜在机会，为企业创造新的增值服务。例如，企业可以利用AI-Native技术提供个性化推荐服务、智能客服系统、自动化生产等，进一步提升客户体验和业务创新。这种创新不仅帮助企业实现业务转型，还能够在行业竞争中脱颖而出，为企业创造更多的增长机会。

3) 开发与部署自动化

AI-Native技术能够推动业务系统的自动化开发与部署进程，减少人工干预，提高开发周期的效率。自动化工具和框架使得开发者能够更加专注于创新，减少重复性工作。机器学习和优化算法帮助自动调整系统配置，支持快速的迭代和部署，提升敏捷性。

4) 运维与优化自动驾驶

通过AI-Native技术，运维流程有望实现高度的自动化和智能化。系统能够自我监控、故障预测并自动修复，减少人工运维的依赖。同时，基于实时数据和机器学习的优化模型，AI能够持续提升系统性能和可靠性，实现真正意义上的自动驾驶运维。

5) 商业模式创新与产业升级

AI-Native技术有望为企业提供创新的商业模式，尤其在服务定制化、智能产品开发及平台化经营方面。AI不仅可以提升产品和服务的附加值，还能够为传统行业带来数字化转型的动力，推动产业的技术升级与结构优化，从而促成新的价值创造和市场竞争力的提升。

2.3 AI Native架构设计方法论

AI-Native架构设计强调将AI技术作为整个系统的核心组件，从数据采集、处理到决策执行的每个环节都与AI紧密结合。与传统的嵌入式AI（Embedding AI）设计不同，AI-Native架构更加注重AI的“无缝融合”，即在架构的每一层都充分发挥AI的能力。设计时，需要根据具体业务场景进行个性化定制，保证技术与需求之间的高度契合。AI-Native架构的设计方法论首先强调数据的重要性。数据是AI系统的基础，通过数据的采集、清洗、存储与处理，AI系统能够在最初阶段便具备足够的信息源来进行学习与决策。其次，架构设计还需要关注智能模型的持续更新与优化能力。在实际应用中，AI模型可能会面临快速变化的市场环境和数据特征，因此系统需要具备自学习、自适应的能力，确保模型能够根据新的数据进行实时调整。

此外，AI-Native架构还强调多层次的技术整合（如图1所示）。在架构的设计中，通常会包括数据层、计算层和应用层等多层次结构，各层之间要形成高度协同，确保整个系统的运行效率和智能化水平。每一层的技术选择要根据具体业务需求来决定，例如数据层使用高效的分布式存储技术，计算层使用大规模并行计算框架，应用层则通过接口与前端系统进行交互。同时，工程部署层面的优化与企业组织的业务形态都应随AI-Native架构的特点进行适配，实现系统智能的可持续性迭代。

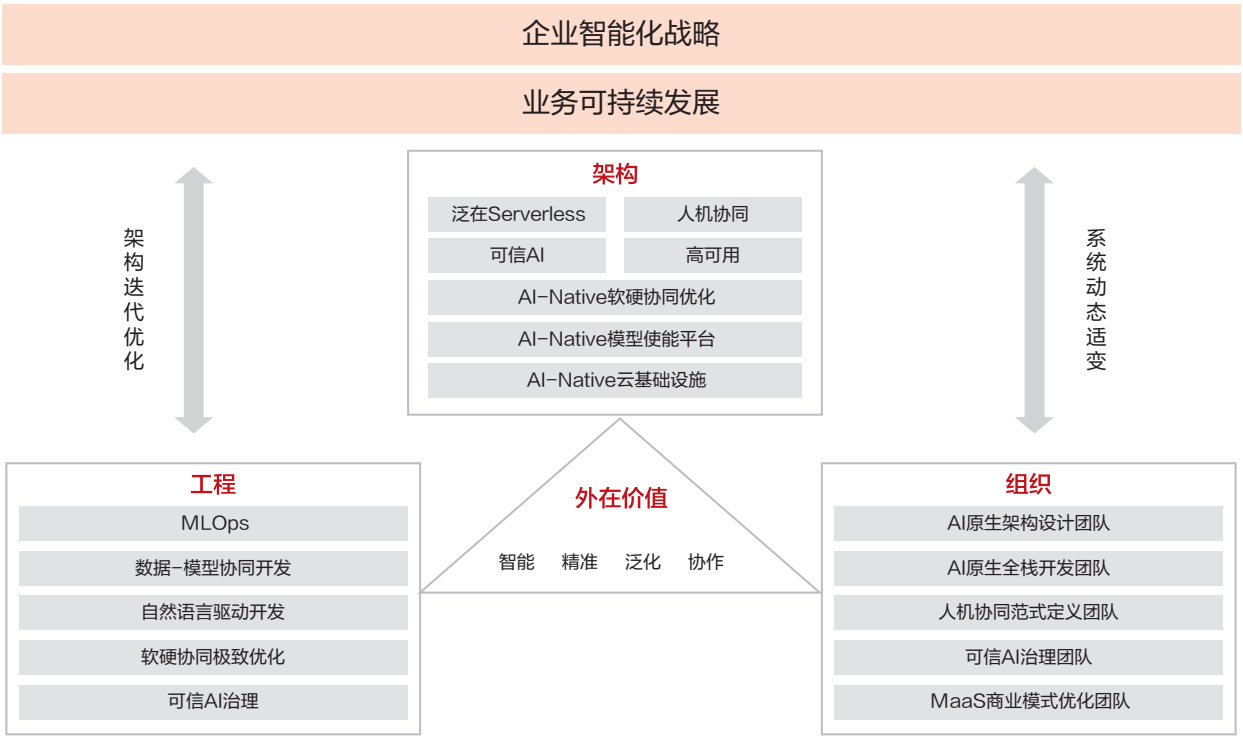


图1 AI-Native系统涉及的架构、工程、组织变化示意图

2.4 Native架构成熟度评估标准

AI原生架构的成熟度可以从多个维度进行评估，包括系统的协作水平、数据治理能力、模型生命周期管理能力、运维自动化程度、系统自进化能力等。根据这些特点列出了如下表所示的成熟度评估标准。

表1 AI-Native架构成熟度评估标准

	Level-0（传统级）	Level-1（入门级）	Level-2（基础级）	Level-3（标准级）	Level-4（发展级）	Level-5（成熟级）
架构	无 AI 架构定义	基础的 AI 参考架构	A 赋能的运营运维及共享的 AI 服务	支持 AI 所需的流数据及分布式计算	完善全面的 AI 架构定义	通过 AI 管理的 AI 架构
协作	AI 功能之间无协同	部分 AI 功能之间通过数据共享协同	部分 AI 功能与基础核心 AI 基础设施平台集成	AI 能力遍布整体架构，同时覆盖 AI 应用、AI 平台及 AI 基础设施	上一层级 AI 系统之间的协作	通过分布式 AI 模型及其智能体应用的广泛协作，实现能力联邦及模型与洞察力共享
数据注入、存储及处理	手动、离线数据管理	自动化的数据收集与在线分析	部分兼容数据资产导入及数据湖架构	全面兼容支持数据资产导入及数据湖架构	全面支持数据湖流水线、数据资产交换网格及零拷贝数据共享	AI 驱动的数据治理及数据资产交换自动化
模型生命周期管理	无专用的模型生命周期管理	手动模型部署	自动模型部署	参考国家地区及行业安全隐私规范的模型适配与数据脱敏，基础 AI 模型的安全可信	自动化模型迁移与升级，增强 AI 模型安全可信	完全自动化的模型生命周期管理及安全
AI驱动的自动化、标准化	私有、非标的日志、告警、性能及配置管理	AI 驱动自动化故障与事件感知，自动化配置与监控	AI 驱动自动化故障定位，性能优化及故障与性能预测	AI 驱动自动化系统修复及抢占式韧性保护	AI 驱动自迭代增强的业务需求管理	AI 驱动的架构设计、详细设计及代码开发测试

03

AI-Native 技术架构

3.1 AI-Native技术总体参考架构

如图2所示，本白皮书将AI-Native技术总体架构分为三层：AI-Native资源层、AI-Native OS层、以及AI-Native应用层。**AI-Native资源层**：构建AI-Native系统的算力基石。作为AI-Native体系的底层支撑，提供适应AI负载的弹性、高性能、异构化资源池，解决传统云架构在算力调度、数据吞吐及网络通信上的瓶颈。**AI-Native OS层**：AI能力的操作系统。作为连接基础设施与应用的“中间件”，提供模型开发、数据治理、基础能力的标准化平台，降低AI应用构建门槛。**AI-Native应用层**：百模千态，赋能业务。面向垂直行业与场景，通过模型调优与业务流程集成，实现AI能力的最终价值闭环。

图2 AI-Native技术总体架构示意图

具体地，**AI-Native资源层**是AI Native时代专门面向多模态大模型、小模型的训练、推理及其智能体应用打造的极致性价比、极致弹性、极致高可用的基础设施，其核心价值体现为在云算力基础设施层为大模型的并行预训练、基于强化学习的后训练、并行推理，以及Agent的任务执行提供最优的计算效率及高可靠保障，并支持能够根据训推任务的实际需求智能调度和分配CPU/GPU/NPU等多元算力以及存储和网络资源，对依据大模型训推及Agent任务的动态变化对上述资源进行动态灵活地弹性伸缩，同时也能支持多租户小模型在云上的高效资源共享与安全隔离。

AI Native资源层在计算架构方面的关键特征：

1 对等算力架构	突破传统主从架构，实现计算节点的对等互联，降低分布式训练中的通信瓶颈。
2 解耦池化	CPU、GPU、TPU、NPU 等算力资源池化，按需组合，提高利用率。
3 Serverless化	支持动态伸缩的算力供给模式，适应AI负载的动态需求。
4 多元算力	异构计算架构（如 GPU/TPU/FPGA）协同优化，匹配不同AI任务的多样化计算特性。
5 分布式边缘计算	边缘节点与中心云协同，实现低延迟推理与数据本地化处理。
6 AI使能弹性调度	基于AI负载预测的智能资源调度，优化算力分配。

在存储架构方面的关键特征：

1 存算分离	计算与存储资源独立扩展，避免存储 I/O 成为性能瓶颈。
2 分布式缓存加速	高频数据就近缓存，减少数据访问延迟。
3 数控加速	存储与计算协同优化（如 RDMA 存储访问），提升数据吞吐效率。
4 多模记忆存储	支持长短期、多模态记忆存储模式，使能Agent实现智能、上下文感知和个性化的交互。

网络架构方面的关键特征：

1 超低时延支撑	如通过RDMAv2/灵衢UB等总线式网络协议实现免CPU介入的节点间超低时延通信，满足CPU与GPU/NPU，以及GPU/NPU与GPU/NPU间频繁的算子执行与运行状态同步需求。
2 超大带宽保障	超节点内通过RDMAv2/灵衢UB无阻塞网络拓扑架构，以及动态拓扑感知的按需网络带宽时空调度满足大模型并行GPU/NPU计算单元之间的参数同步需求，最大限度避免并发集合通信冲突带来的性能影响。
3 应用驱动的极简云网络	对传统数据通信协议栈，特别通过数据中心内应用网格、虚拟叠加网络以及物理承载网络的深度融合，以及跨数据中心P2P模式的网络路由、传输层协议进行基于控制承载分离模式的全面简化与重构，从而大幅提升面向AI Native时代的网络交互效率。

AI-Native OS层是连接底层算力与上层应用的智能中枢，提供模型开发、训练、部署的全栈支持，并构建AI模型的训推提供最为关键的数据飞轮体系支撑。此外，L0级基础大模型也面向提供通用AI能力，作为行业模型的基座。按模型赛道分为NLP大模型、CV大模型、预测大模型、科学计算大模型、多模态大模型（包括多模态理解大模型、多模态生成大模型）等。每一类L0基础大模型通常包含多种候选，如NLP领域包括DeepSeek V3、盘古NLP系列、Llama系列等，在实践中可以设计一个模型调度与路由机制层，自适应加载不同的L0大模型。各L0大模型均由大量通用、领域数据训练得到，可以在部分行业场景中开箱即用，也可以配合工具链优化其部分能力，从而得到面向特定行业的L1大模型或面向特定场景的L2大模型，进而使能上层百模千态的AI应用。

实践中，在构建AI Native系统/应用时，实现算力与模型性能的极致协同需要软硬协同的深度耦合，即：**AI-Native软硬协同优化**。例如，通过稀疏MoE架构设计，在硬件层面实现异构算力的动态调度，使专家模块的并行计算与硬件资源的亲和性达到最优匹配；混合精度技术则通过FP32/FP16等精度的智能切换，在保证模型精度的前提下，有效提升硬件计算单元的吞吐量；而基于流水线的通信隐藏技术，通过计算与数据传输的时空重叠，充分降低跨节点的通信开销。这种从算力架构到模型结构的全栈优化，最终形成“算力-算法-数据”三位一体的协同增强效应，为AI-Native系统构建出性能更高的技术基座。

AI-Native应用层采用分层设计，包含行业大模型（L1）、场景大模型（L2），以及基于二者构建的百模千态应用。这些应用可分为三类：AIGC类（如图像生成、视频生成、文本生成等任务）、Agent类（如通用智能体Manus、金融分析Agent等）、以及AI-Native云服务（如华为云Versatile等）。在实际系统构建中，开发者可能组合调用L0/L1/L2模型，或直接使用上层应用。若涉及Agent开发或多Agent协同，需通过MCP协议（Multi-agent Control Protocol）实现对外部知识库、工具链的调用（如检索增强生成/RAG、API工具调用），以及A2A协议（Agent-to-Agent Protocol）来支持Agent身份认证、Agent间状态同步（如任务上下文共享）等。例如，政务场景中，一个基于L1政务大模型的公文审核Agent可能通过MCP协议调用法律条文数据库，同时通过A2A协议与多个部门审批Agent协同工作，实现全流程自动化。

以华为云自身实践为例，基于华为全栈云服务API/SDK文档、生态部上云最佳实践、海量上云解决方案案例库、外部互联网IT通识（IT解决方案通用知识与案例）、行业领域特定知识库等语料库，微调盘古L0级基础大模型，进一步结合具体的业务场景和流程，打造出安全云脑、CodeArts、云运维、云数据库等一系列垂域模型，赋能相关场景/应用的服务智能化转型。

3.2 AI-Native资源层关键技术解析

3.2.1 对等计算、解耦池化的多元算力AI超节点

» 趋势和需求

AI算力资源发展至今，从传统的CPU到GPU，再到百家齐鸣的NPU、TPU、DPU等等，AI云计算已经进入了一个高速发展的XPU时代。在AI算力业务蓬勃发展的时代背景下，AI算力诉求急剧膨胀，从最开始的单机单卡、单机多卡，到现在的千卡、万卡集群，这也引出了一系列的问题和挑战：

集群规模快速膨胀，AI资源管理复杂度上升。随着AI产品的大众化、规模化，搭载商业级算力芯片的大规模算力集群，成为了各个科技型企业的必备武器，AI算力集群规模也日益膨胀，这就带了不可避免的问题：如何能更高效的管理成千上万的AI算力资源。

AI芯片种类繁多，对于AI资源管理的可扩展性有了更高要求。无论是现今一家独大的英伟达，还是厚积薄发的华为、谷歌、AMD，都在推出AI场景算力芯片，例如英伟达的GPU、华为的昇腾NPU及谷歌的TPU。AI算力云厂商或是AI型企业，面对各家算力厂商迥异的架构，也急需有一套可扩展性更好的AI资源管理架构。如下图所示，异构资源通过池化方式，为上层训推等任务提供算力支持。

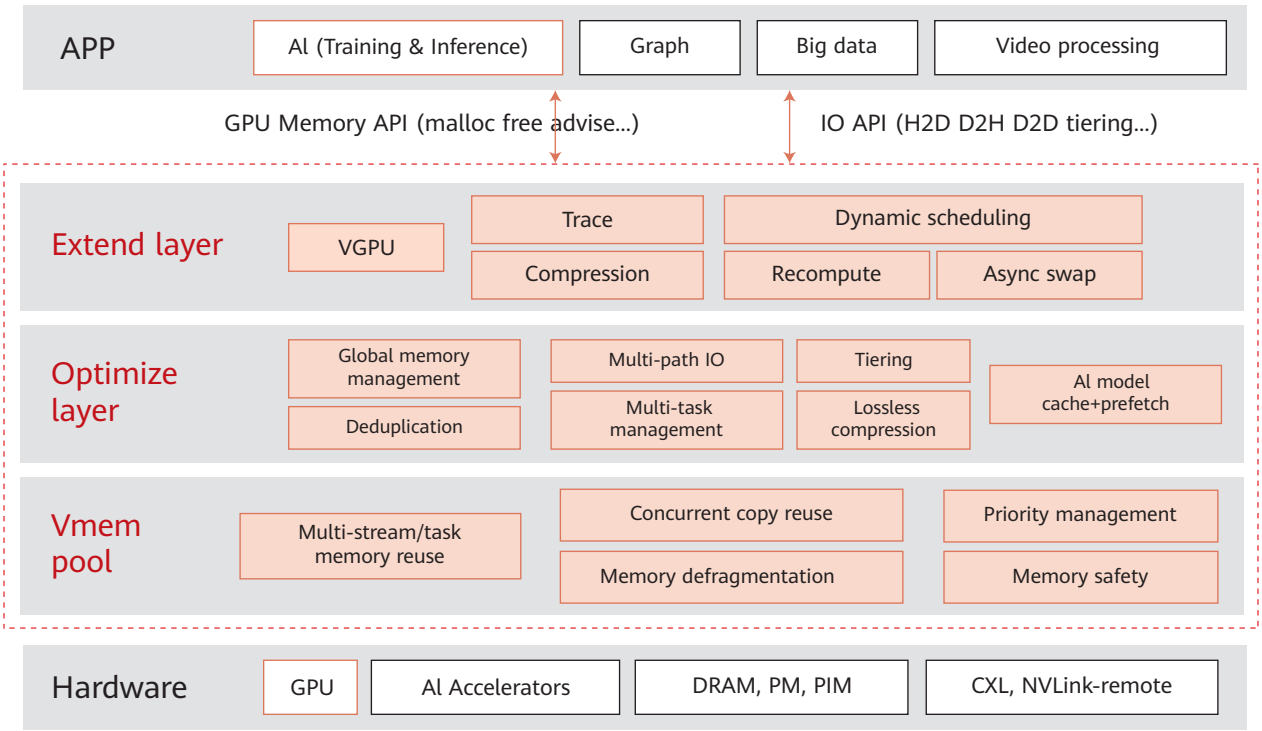


图3 异构资源池化架构图

参数面网络等新型AI资源，对于AI资源管理提出了新的挑战。大模型、自动驾驶、AIGC的横空出世，大规模的算力参数面互访网络成为了必需品，参数面网络提供的超高带宽，发展出了计算机超节点架构，计算机超节点是一个由多个和多种计算(CPU/NPU)，内存，IO设备等计算机资源单元，高速互联紧耦合在一起的集群计算系统，是生成式AI时代的产物。区别于传统以服务器中心松耦合架构，超节点是去中心化的紧耦合架构。随着技术的进一步演进，未来超节点内所有服务器的设备可做到灵活组合成为各种算力单元，也可被称为矩阵式算力。为了能够有效利用超节点内的资源，相关联的算力参数面网络设备及其拓扑的管理，也就成为了AI算力资源管理的新课题。

» 关键特征与相关技术

面对问题和挑战，作为AI云原生基础设施资源底座，kubernetes构建了面向超节点架构的整套资源管理方案。

虽然计算机超节点的High-Speed Link高速互联能够提供比传统互联更高的带宽，但单路径带宽仍无法匹配计算单元的吞吐，基础设施层通过构建全局多路径I/O加速技术，大幅提升了节点内与节点间I/O性能。

为匹配AI行业所需的庞大算力需求，基础设施硬件从主从架构逐步演进至对等架构，传统的资源管理模型不再适用，需要构建面向对等架构的资源管理模型，实现资源的高效管理与合理配置。

传统资源管理模型的基本算力单元为单台服务器，服务器模型内包含各种设备（CPU，内存以及I/O设备等），资源池模型由服务器模型聚合而成，其资源分配也是以服务器为基本粒度，云化场景下的云服务器也仅是设备数量存在差异，其基本建模均保持一致。超节点为去中心化的架构，虽然物理设备仍依托于服务器之上，但超节点内配备有超高速互联网络，其内所有设备均可以灵活组合成不同的算力单元，超节点架构基本算力单元不再是单台服务器，传统资源管理模型已不再适用。面向超节点架构，Google的TPU服务构建的层次化的资源管理模型，是业界当前比较成熟的解决方案。

1) **超节点资源管理模型与资源切片：**超节点资源管理模型包含三个基本算力单元模型：XPU、CPU和内存，其他设备均建模为附属模型。在资源管理模型中将基本模型又被抽象为资源节点Node，超节点的高速互联被抽象为连接资源节点之间边Edge，一个超节点被抽象为一个SuperPoD，多个SuperPoD组成一个集群Cluster，资源池就是集群的聚合。

   SuperPoD的资源分配模型是XPU、CPU和内存的组合，称为超节点资源切片slice。其中XPU的资源分配粒度为设备，CPU为CPU Core，内存为容量。Edge作为资源组合的约束，对资源的组合形式进行限制。比如客户申请一个64XPU，320CPU Core，1024GB内存的slice，超节点资源调度器不仅要调度足量的XPU、CPU和内存资源，还要通过图匹配算法确保被调度的资源节点之间存在直连Edge。基本算力单元之外的设备不参与资源调度过程，而是通过规格预定义的方式进行管理，在AI场景下这些设备的分配量一般与XPU资源量锚定，按照不同的XPU请求量划分为若干档位。

2) **超节点资源拓扑感知：**AI业务场景下所需的通信量非常大，其通信算法都会根据基础设施网络拓扑进行编排优化，以达到充分利用网络带宽的目的。为了有效利用超节点的高速互联网络，客户也需要感知到超节点内部的拓扑结构来优化通信算法。然而算力服务提供商出于安全和保密方面的考虑，一般不会对客户暴露物理信息，而是通过抽象方式隐藏物理信息。AWS提供了一套网络拓扑的抽象建模思路能够在满足通信算法优化需求的同时隐藏物理信息。超节点资源拓扑感知模型将不同的网络设备抽象为虚拟的网络节点NN（network node），并为每一个NN进行逻辑编号，如NN001，NN002。客户在查询超节点slice的设备拓扑时，接口会返回每一个设备所属的每个层级的NN，客户可以根据NN的逻辑编号是否相同来确定设备间高速互联的拓扑结构。

3) **超节点资源高可用：**高可用能力是大规模集群系统必须具备的基本能力，基础设施层的高可用能力之一是故障设备替换。故障设备替换指的当客户正在使用的设备出现故障时，使用一个正常设备将其替换掉，帮助客户快速恢复业务。在超节点架构下，由于超节点内的设备之间具备高速互联网络，所以可用于替换的设备必须在超节点内部，不能跨超节点进行设备替换。在超节点架构下执行故障设备替换时，资源管理平台会约束调度系统的调度范围不能超出设备所在的超节点。此外，由于超节点规模有限，为了确保超节点内存在可用于替换的设备，资源管理平台会在每个超节点内预留部分设备作为保底手段。在故障替换时会优先选择非预留的空闲设备，在非预留空闲设备不满足替换需求时才会动用预留资源。在某个预留设备被使用后，预留设备池的容量随之减少，资源管理平台会周期性的扫描超节点内设备使用状态，若存在被释放的设备则将其加入预留池，以实现预留池容量的轮转。同时，资源你管理平台也会通知运维人员及时维修故障设备。在AI场景下，为了与Checkpiont机制相配合，资源管理平台会对外暴露设备替换接口。AI作业管理平台在保存好现场后调用此接口进行故障设备替换，替换成功后再通过读取checkpoint恢复业务。

除设备故障外，网络断连也是典型的故障场景，超节点资源管理平台采用借轨通信的方案解决此类问题。借轨通信是指在设备A与C的当前互联路径中断的情况下，由于设备A和C仍然与设备B保持通信连接，设备A可选择从设备B跳转的方式与设备C实现通信。跳转节点通过路径规格算法进行优选。

3.2.2 软硬解耦、细粒度资源调度

随着大模型高速发展的背景下，引发AI算力需求指数级增长。然而当前AI大模型算力服务面临诸多问题，如集群资源利用率低、资源分配粒度僵化、异构资源管理困难等。本章节将展开业界全域调度和弹性细粒度调度趋势、挑战及相关技术。

» 趋势与需求

在全球数字化转型和DeepSeek等大模型高速发展的背景下，AI算力需求指数级增长。据IDC预测，2023-2030年全球IDC市场将保持22%的年复合增长率，而中国智能算力增长更为迅速，预计2025年算力规模会突破千亿级。其中，昇腾在中国智能算力占有重要的地位，昇腾云服务提供了高性价比的AI算力，包括910A/B/C等多种NPU硬件，并提供全链路云化工具链，支持高效迁移，全栈垂直优化，以及模型/算法高效运行，使能“百模千态”应用快速落地。

随着Scaling Law的持续演进，模型参数和计算量增长迅猛，开发者对云上AI算力的成本和可靠性等诉求也越来越强烈。具体来看，AI云服务当前面临着问题和挑战：

1) **资源分配力度过大，导致资源利用率低用卡成本高。**当前GPU/NPU算力分配颗粒度过大，无法准确匹配多样化AI任务的资源需求。以固定资源规格匹配多样化的算力需求，造成大量资源碎片，资源利用率偏低。

2) **潮汐效应，导致资源浪费严重。**推理任务白天和晚上调用量差距明显（白天调用量是晚上的10倍），如按峰值预留资源，则低谷资源利用率极低，造成资源浪费。

3) **大模型分布式集群故障频发，训推可靠性低。**大规模并行+耦合计算导致“故障爆炸半径放大效应”，难以高效稳定支持超大模型训推。此外，在训练阶段，CKPT频率设定挑战大，高频浪费资源，低频回滚耗时，导致故障恢复慢。

» 关键特征与相关技术

面对这些问题和挑战，业界共识是通过细粒度调度和软硬解耦，来构建更便宜、更可靠的算力服务，使能AI创新，具体特征和技术如下：

1) **细粒度切分、量体裁衣，提升资源利用率。**通过细粒度切分和分配，打破算力资源规格枷锁，支持GPU/NPU、内存/显存、存储IO，以及网络吞吐等维度的任意比例组合，基于对云上业务负载的高精度、细颗粒资源画像，自动推荐与业务负载需求匹配的最优算力资源配比。如下图所示，细粒度切分与分配相对于整卡分配，节省25%算力、63%容量、13%访存带宽及25%网络带宽。

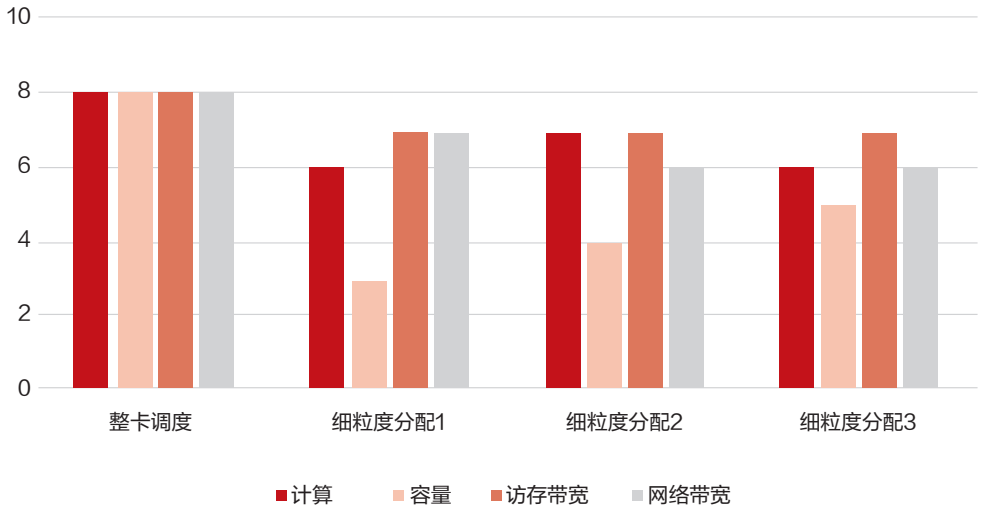


图4 Llama 8B 910B3 2k/2k，满足SLA约束的NPU切分方法方案

2) **智能弹性伸缩+训推混部。**由于AI应用潮汐效应明显，导致资源浪费，可通过时序预测AI模型，提前预判业务负载波动，实现资源供给与需求曲线的精准贴合。并通过训推混部方式，将调用量低谷时多余的推理资源腾挪给训练使用，在保证SLA前提下使得AI算力成本进一步下降，如下图所示：

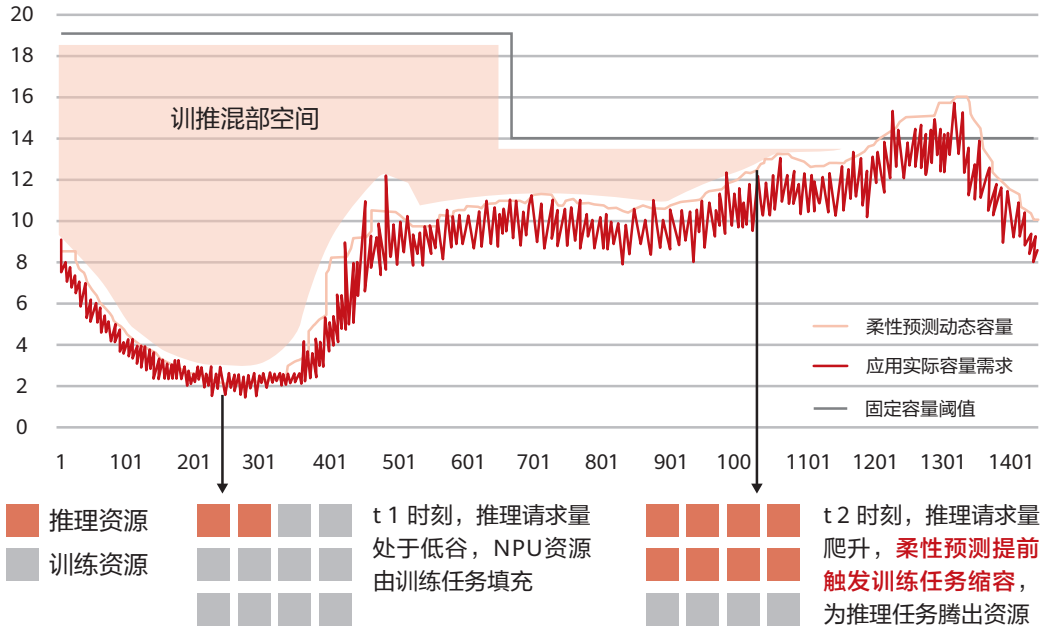


图5 AI算力资源预测与训推混部

3) **软硬解耦屏蔽故障，系统快照故障快速恢复。**可通过NPU虚拟化技术，实现训推任务与NPU硬件的解耦，通过“软件定义NPU”屏蔽NPU硬件故障对训推任务的影响。此外，可通过构建系统级快照，出现NPU故障时，直接从剩余NPU节点中获取完整任务状态，并重新拉起，将训练损失缩减为单步内，提升系统的可靠性。

3.2.3 存算分离、极致IO吞吐的AI原生云存储

» 趋势与需求

基于大模型的生成式AI技术的重大突破推动了人工智能的应用范围从传统的分类任务扩展至广泛的生成任务，引发了AI应用的爆发性增长，并引领IT产业迈入全新的“AI时代”。随着AI产业的迅猛发展，云计算基础设施也在从以通用算力为核心向以智能算力为核心转变。在这种新型云计算基础设施中，数据的“算力”和“存力”是相辅相成的。一方面，强大的数据算力（包括GPU、NPU等算力单元）需要充足的数据存力（如显存、DRAM、SSD等存储单元）来保证数据处理的连续性和稳定性；另一方面，高效的数据存力也需要数据算力的支持，以便对存储的数据进行有效处理和利用。数据的算力和存力之间存在着紧密的联系和相互依赖。在全局视角下，要提升端到端的效率，“算力”跑的快，“存力”也要跟上，算力与数据存力一起系统化地构成了AI 算力基础设施。尽管云数据中心在智能算力方面取得了显著进步，但是在存力方面的不足已成为制约效率的关键瓶颈。

1) AI训练主要有以下两个存力问题

i. CheckPoint 保存与恢复慢导致GPU/NPU算力利用率降低

大模型训练AI集群故障概率高，故障影响大，故障发生后任务恢复耗时长，浪费大量AI算力和时间。AI算力集群可用度和算力资源利用率问题是AI集群使用者和供应者共同关注的问题，集群的可用度直接关系到AI训练任务能否在预期的时间内完成，而可用度和算力资源利用率对企业内的AI基础设施部门或公有云厂商则意味着服务SLO能否达成，能否通过压低AI集群的资源成本取得盈利。以Meta的OPT-17B训练为例，理论上在1000个80G A100上训练3000亿个单词，需要33天，实际训练却使用了90天，期间出现了112次故障。如下图所示，集群卡数规格越大，平均故障间隔时间MTBF越短，而故障恢复时间MTTR的快慢则直接影响到集群的可用度和算力资源利用率，严重时集群算力资源利用率只能达到33%，导致2/3的算力被浪费。

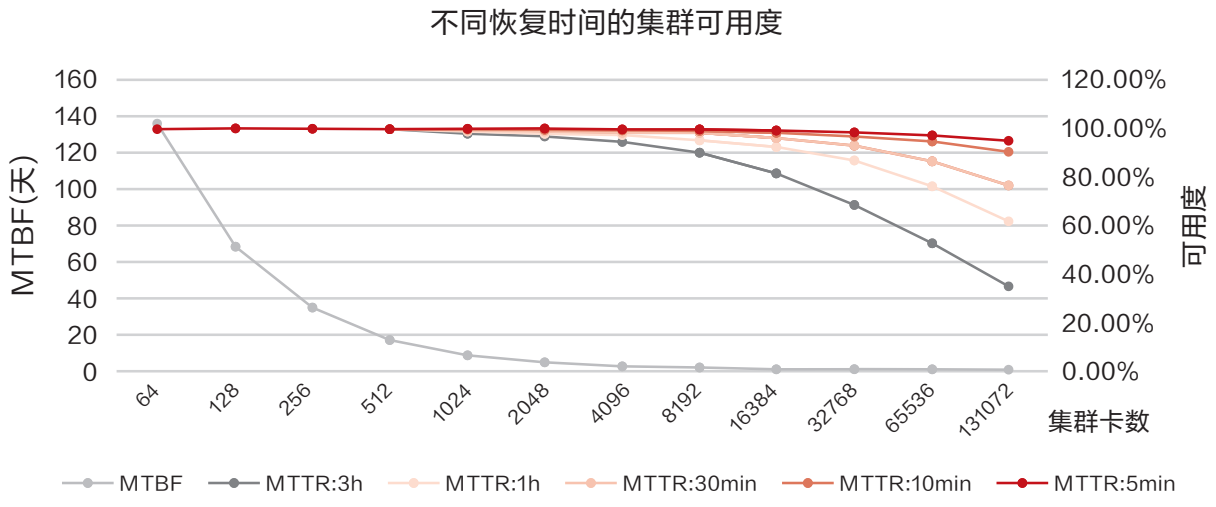


图6 集群可用度

训练任务检查点CheckPoint是深度学习框架中的容错方法。检查点CheckPoint通过在给定时间定期保存完整模型状态的快照来帮助缓解训练的模型状态丢失问题。如果发生故障，可以使用之前保存的CheckPoint快照将模型重建到快照时的状态，以从该点恢复训练。但是，根据CheckPoint检查点保存频率，通常会导致几个小时的计算时间损失。此外，保存和恢复CheckPoint过程本身会产生大量开销，恢复时所有节点都需要并发读取CheckPoint，千亿大模型TB级大小的CheckPoint文件保存和恢复通常会成为训练过程中的瓶颈，CheckPoint保存和恢复过程中会长时间中断训练任务，浪费大量算力和时间，考虑到大模型使用的GPU/NPU规模，以1万卡为例，故障损失将会是数万个卡的时间。下图说明了训练过程中CheckPoint保存和故障恢复时的时间和算力开销。

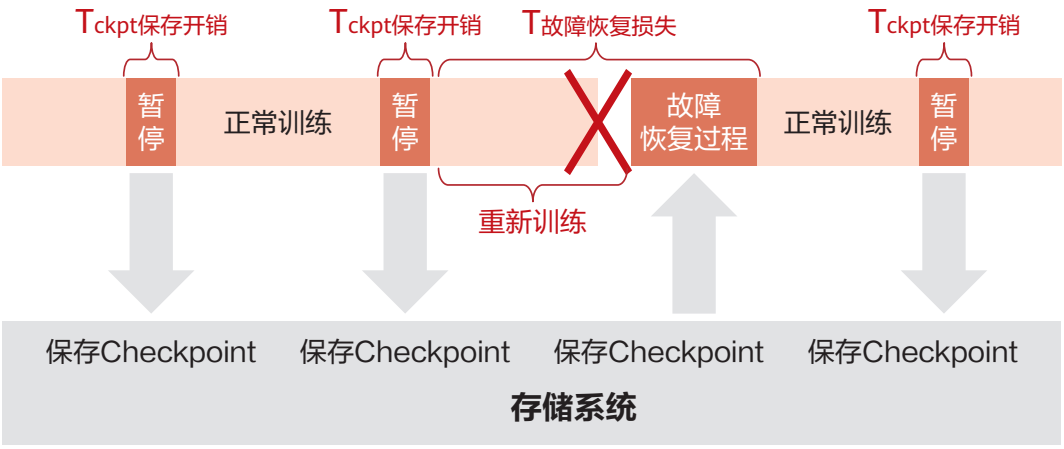


图7 训练过程中CheckPoint保存和故障恢复时的时间和算力开销

另外，大模型的参数数量呈指数级增长，导致模型大小急剧增加，模型参数量越大，CheckPoint文件越大，由于CheckPoint中除了包含模型参数权重信息，还包含优化器、配置等训练任务信息，膨胀系数一般按6倍计算。例如，GPT-4的参数数量为1.76万亿。若使用FP16格式存储模型参数，GPT-4模型参数约为3.52TB，在模型训练过程中，模型的CheckPoint大小是21.12TB。按1min完成CheckPoint保存，5min完成CheckPoint恢复，数据并行度DP为30，则对存储系统的写带宽需求352GBps/s，读带宽需求为2112GBps/s。总之，大模型CheckPoint检查点管理涉及繁重的存储和作业恢复时间开销，频繁的CheckPoint检查点保存，加上从最近可用的CheckPoint检查点快速恢复训练作业，成为一项巨大的挑战。其中，CheckPoint保存与恢复过程对存储系统读写带宽要求的计算公式参考如下：

存储写带宽 = CheckPoint大小 / 保存时间

存储读带宽 = CheckPoint大小 * 数据并行度DP / 恢复时间

CheckPoint大小 ≈ 模型参数量 * 参数占字节数 * 膨胀系数

ii. 海量训练数据加载慢导致训练任务变慢

CV/多模态/自动驾驶等训练任务场景涉及PB级训练数据，数据集读取慢导致GPU/NPU算力出现空闲，训练任务变慢。随着企业使用 GPU/NPU 算力规模越来越多，底层存储的IO 已经跟不上计算能力，企业希望存储系统能提供高吞吐的数据访问能力，充分发挥GPU/NPU的计算性能，包括训练数据的读取加速，减少上层算力对存储 I/O 的等待。以CV类大模型场景每台AI训练服务器图片处理速度10000个/s，平均每张图片大小约为200KB，按照千卡规模128台计算节点并发训练，读带宽性能需求为10000*200*128 =244GBps。总带宽需求随计算节点规模线性增长，算力规模越大，存力需求越高。

2) AI 推理业务对存储的主要挑战

面向AI推理业务，存力面临的痛点主要表现在三个方面: 持久化存储性能不足、DRAM利用率低、以及AI内存墙问题。

i. 持久化存储性能不足

近年来，大模型的参数数量呈指数级增长，导致模型大小急剧增加。在模型推理场景中，AI加速器需要将模型文件加载到其显存中进行推理，特别是在推理集群发生故障恢复、推理业务高峰期发生弹性扩容的时候，各AI加速器节点需要从共享存储中并发快速完成模型文件加载。在Serverless推理场景中，AI加速器还需频繁切换不同的模型以满足不同用户的推理任务需求，这种模型切换的时延需求通常在秒级别，对存储性能提出极大的挑战。以38B参数量的模型文件为例，模型文件大小约为80GB，单个模型文件被拆分加载到1机8卡进行分布式推理，按千卡规模推理集群分钟级完成模型文件并发加载，对存储的带宽需求为: 80*1000/8/1/60=166GBps，由于存储数据量相对较小，对存储性能密度提出极大的挑战。同样的，推理集群规模越大，对存力要求越高。

ii. DRAM 利用率低

当前的AI集群不仅包含AI加速器，还配备了大量的DRAM内存资源。例如，一台华为AI服务器配置了8张NPU卡和1.5TB的DRAM，而NVIDIA GH200服务器中每张GPU卡则配备了512GB的DRAM。然而，在运行流行的大语言模型训练和推理任务时，这些DRAM资源的利用率却非常低。一项研究论文分析了两个较大规模的GPU集群的资源利用率情况，结果显示其中一个集群在90%的使用时间内DRAM利用率低于25%，另一个集群在76%的使用时间内DRAM利用率同样低于25%，且在几乎整个使用过程中，两个集群的DRAM利用率均未超过50%。AI集群中DRAM利用率低的主要原因在于，AI服务器上的DRAM资源通常是按照各种负载场景的最大需求进行配置的，以确保能够运行所有类型的负载。这种配置策略导致在某些特定负载下DRAM利用率较高，而在大多数其它负载下DRAM利用率则较低。由于目前AI集群主要由LLM负载主导，因此DRAM的整体利用率普遍偏低。

iii. AI 内存墙

AI内存墙主要包括内存容量墙和内存带宽墙两个方面。在内存容量墙方面，AI加速器的显存容量增长速度远远落后于大模型存储需求的增长速度。如图8所示，典型的Transformer大模型的参数量每两年以240倍的速度增长，而业界典型的商用AI加速器的内存容量仅每两年翻两倍。这种大模型参数量与AI加速器显存容量增长速度之间的巨大差距，意味着训练和推理一个模型需要更多的AI加速器，这将显著增加AI训练和推理的成本。此外，增加AI加速器数量的主要目的是为了让大模型能够存储在AI加速器的显存中，这通常会导致AI算力的利用率低下。

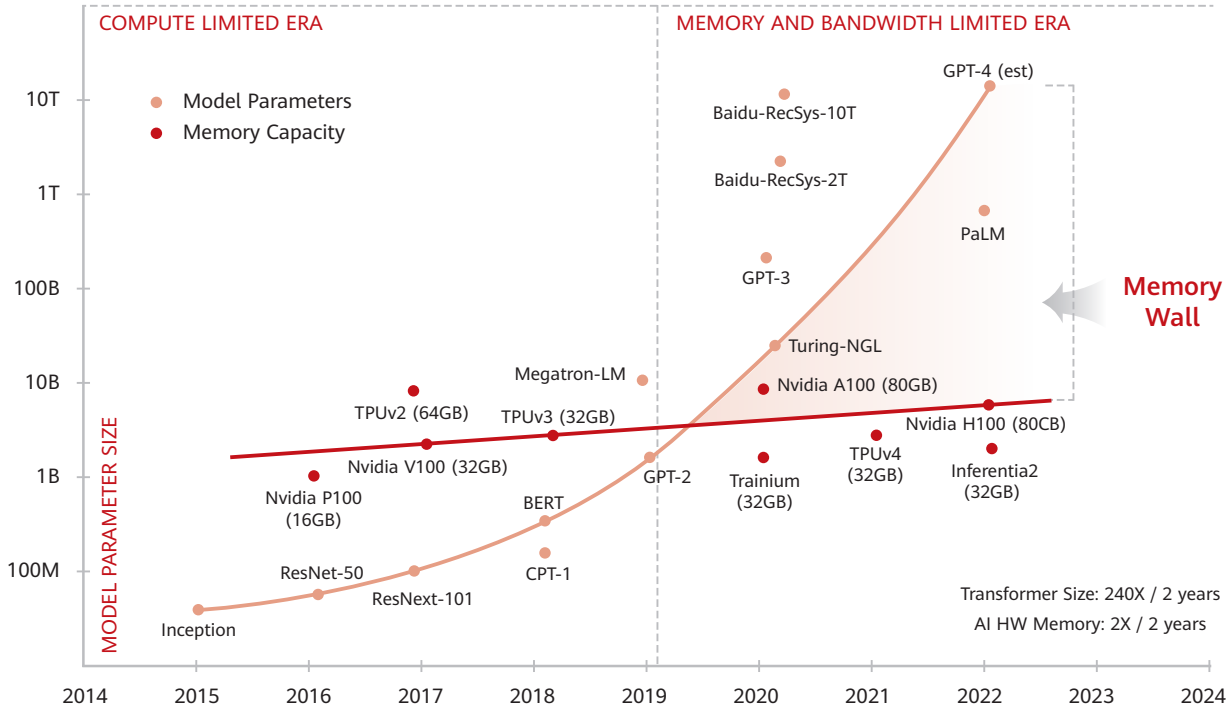


图8 SOTA 模型的参数量增长趋势和AI硬件显存内存容量增长趋势

来源: Amir Gholami, Zhewei Yao, Sehoon Kim, Coleman Hooper, Michael W. Mahoney, and Kurt Keutzer. "Ai and memory wall." IEEE Micro (2024)

在内存带宽墙方面，AI加速器的显存带宽的增长速度远低于其算力的增长速度。如图 9所示，过去20年间，单个AI加速器的峰值计算能力增长了9万倍，而内存访问带宽仅提高了30倍。这是因为提升硬件算力的工艺相对容易，而增加内存带宽的硬件工艺则难度较大。AI加速器内存带宽与算力增长速度之间的巨大差距，意味着在进行AI计算时，往往需要等待数据从内存中读取，这导致算力的利用率降低。

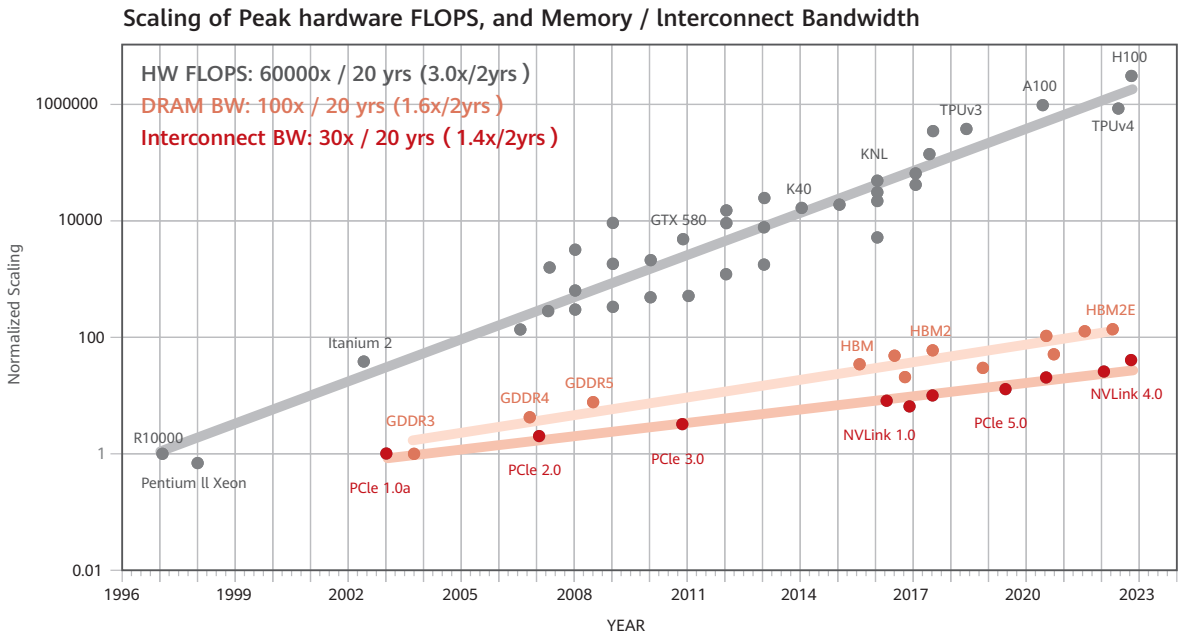


图9 AI加速器的计算能力、内存带宽和互联带宽的增长趋势

来源：Amir Gholami, Zhewei Yao, Sehoon Kim, Coleman Hooper, Michael W. Mahoney, and Kurt Keutzer. "Ai and memory wall." IEEE Micro (2024)

» 关键特征与相关技术

1) 数据联动技术，解决AI训练数据加载慢问题

随着云上对象存储成本的逐渐降低，越来越多的企业利用对象存储保存大量数据并构建数据湖。由于对象存储的性能和生态接口无法满足AI训练数据快速加载的需求，业界一般使用文件系统作为高性能缓存层，而对象存储则作为统一的数据底座，存储大量冷数据，以减少存储成本。用户通过指定高性能文件存储文件系统内的目录与对象存储桶进行关联，然后通过创建数据导入导出任务实现数据同步。当AI训练任务开始前，可以将对象存储数据湖中的AI训练集数据先高速预热到高性能文件缓存加速层中，以实现训练时数据集高速读取，避免AI芯片因存储I/O等待产生空闲，提升AI芯片利用率。

业界主流的云计算服务提供商大多也采用此方案，如AWS的Amazon FSx for Lustre是一款高性能文件系统，它适用于需要高吞吐量和低延迟的AI工作负载，当创建FSx for Lustre文件系统时，可以指定一个S3 bucket和文件系统关联。这样即可通过文件系统透明的访问S3 bucket中的文件和目录。当访问文件数据时，可以实时透明的把数据从S3移动到FSx for Lustre文件系统中，也可以通过命令把数据写回到Amazon S3 bucket中。S3 bucket 用于长期持久化存储大量数据，以减少存储成本，FSx for Lustre更多的用于AI数据加速访问场景。

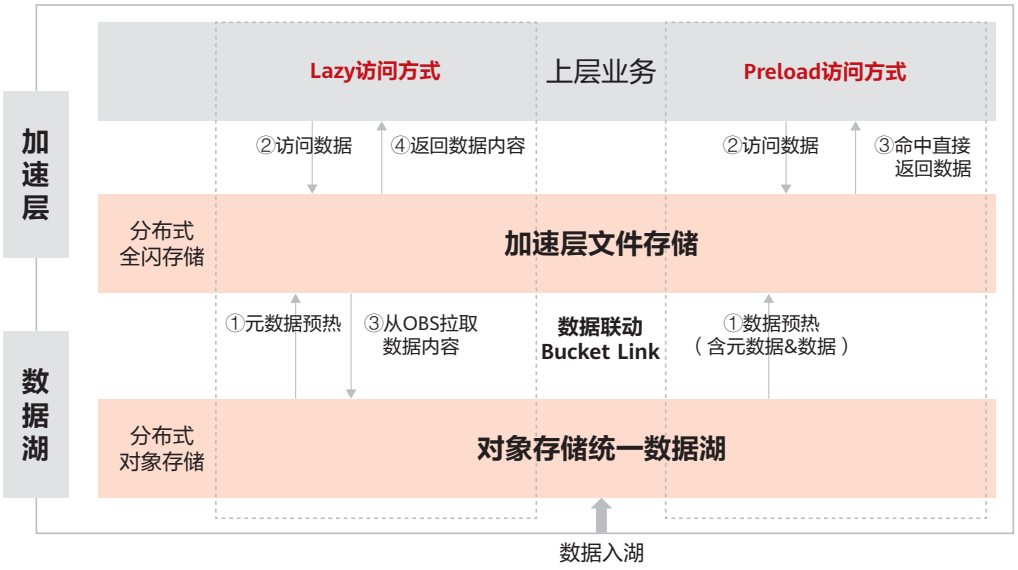


图10 AI模型训练场景

2) 缓存加速技术，满足CKPT快速保存及恢复需求

业界主流的方法是使用训练任务检查点CheckPoint应对AI训练集群故障问题，检查点CheckPoint通过在给定时间定期保存完整模型状态的快照来帮助缓解训练的模型状态丢失问题。如果发生故障，可以使用之前保存的CheckPoint快照将模型重建到快照时的状态，以从该点恢复训练。但是，保存和恢复CheckPoint过程本身会产生大量开销，恢复时所有节点都需要并发读取CheckPoint，千亿大模型TB级大小的CheckPoint文件保存和恢复通常会成为训练过程中的瓶颈。主流的云计算服务提供商普遍基于自研的高性能文件系统来加速CheckPoint检查点的读写请求。具体解决方案是基于高性能文件缓存加速层存储提供L1服务端缓存，客户端内存缓存提供L2内存缓存，满足CheckPoint快速保存及故障时的CheckPoint快速恢复。头部厂商主流是基于文件存储缓存层提供Tbps级高速访问能力，如Meta的Tectonic文件系统、AWS的FSx for Lustre、DeepSeek的3FS文件系统等。

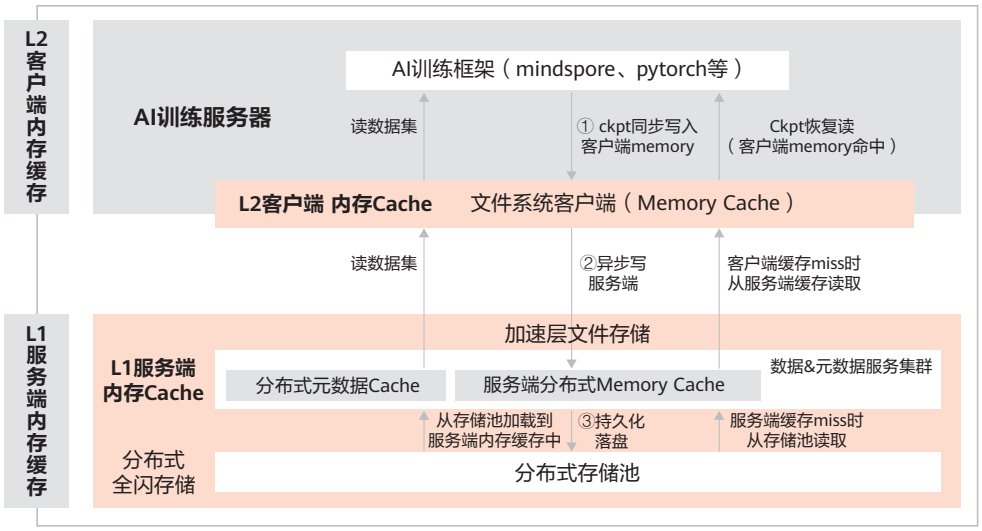


图11 模型训练基于缓存加速CKPT读写场景

3) 计算-内存-存储三层极致分离架构

在AI推理过程中，Transformer模型接收用户的问题输入，并通过迭代方式生成相应的回答。每个Transformer层由自注意力模块和前馈网络模块组成。在自注意力模块中，上下文词元（token）与模型参数结合，生成中间数据K（键）和V（值），并进行注意力计算。为避免在迭代生成过程中重复计算KV，业界主流方案是把生成的KV中间数据被存储在AI加速器的显存内存中，形成KV缓存。每个词元的KV缓存大小取决于模型的维度、层数以及数据精度，计算公式为：单个词元的KV缓存大小 = 模型维度 * 模型层数 * 数据精度 * 2。例如，GPT3模型的数据维度和层数分别为12288和96，在双字节精度下，单个词元的KV缓存大小为12288 * 96 * 2 * 2字节 = 4.5MB。

在推理过程中，每个推理请求所需的KV缓存大小与上下文长度成线性关系。例如，在GPT3模型的推理中，长度为2048的上下文将占用4.5MB * 2048 = 10GB的AI加速器显存内存空间。然而，AI加速器通常只能提供几十GB的显存容量，其中一部分用于存储模型参数，仅剩余有效的空间用于KV缓存。例如，使用8张64GB的AI加速器部署GPT3模型，系统显存总容量为8 * 64GB = 512GB，其中350GB用于模型参数，剩余162GB仅能支持162GB/10GB = 16个2048上下文长度的推理请求缓存KV值。因此，AI加速器能够同时处理的请求数量受限于显存内存容量。综上所述，Transformer模型推理中存在严重的显存内存墙问题。为了解决AI推理中存在内存墙问题，业界可将传统的“计算-存储”分离的两层架构升级为“计算-内存-存储”分离的三层架构，其中新增的“内存层”即为弹性内存存储层。这种极致存算分离的基础设施架构具有高资源弹性、高资源利用率和高性能等优势，能够有效解决上述存力痛点。

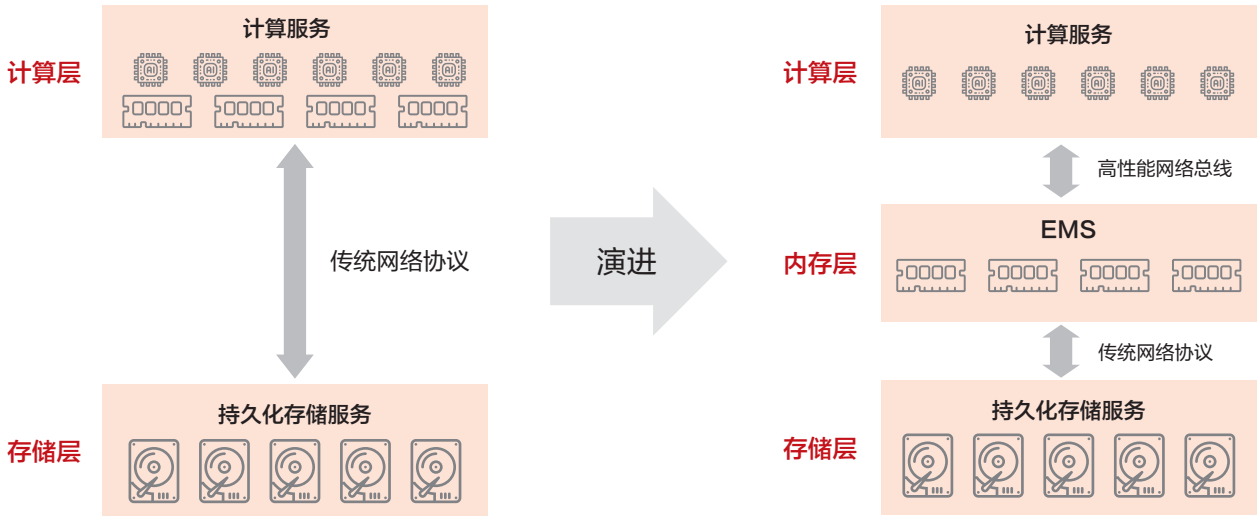


图12 “计算-内存-存储”分离的新型三层云架构演进

3.2.4 无阻塞、确定性低时延的AI原生云网络

» 趋势和需求

1) AI原生云数据中心网络发展趋势和需求

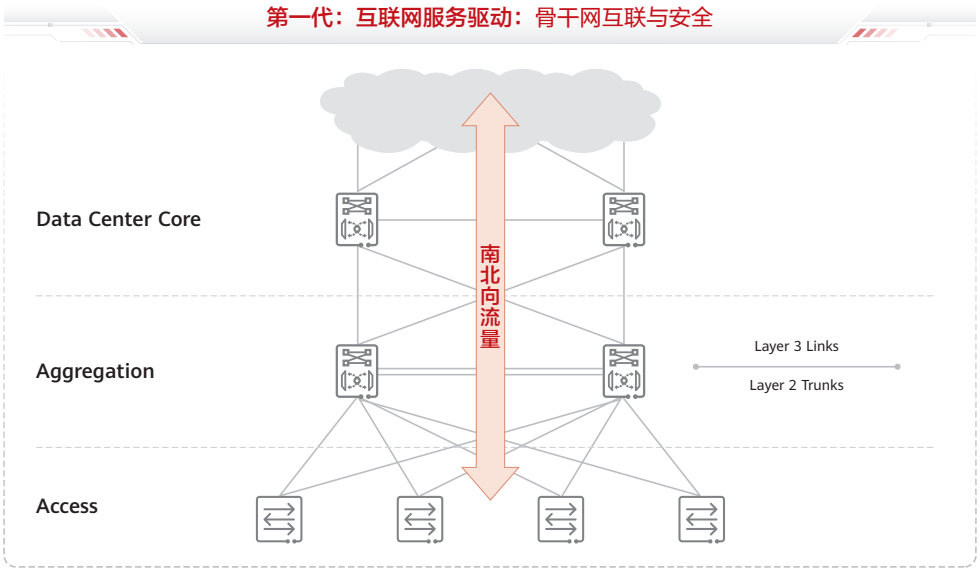


图13 数据中心网络架构1.0

数据中心网络架构1.0时代：南北向流量为主，以服务为核心的互联网服务驱动架构，通过互联网提供的各类服务来构建、扩展和整合应用系统。1.0网络架构中业务对网络QoS需求为：接入带宽：1Gbps ~10Gbps，网络时延：10ms~100ms，丢包率：0.1%~2%。

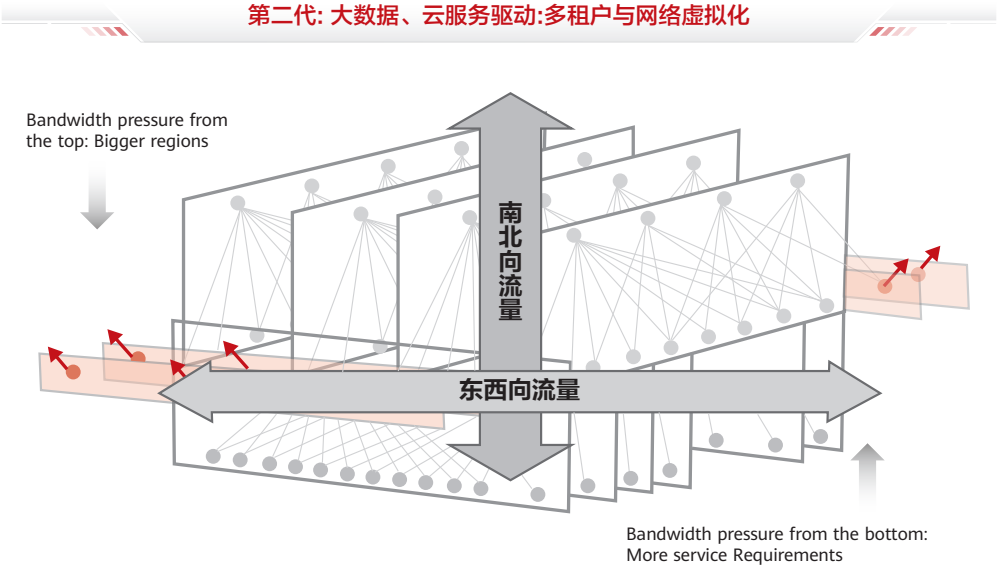


图14 数据中心网络架构2.0

数据中心网络架构2.0时代：相比1.0时代，由南北向流量为主变为兼顾南北向和东西向流量，由互联网服务驱动架构变为大数据、云服务驱动。云服务驱动网络指网络基础设施（服务器、存储、网络设备等）由云服务商提供，通过云原生技术实现灵活管理和自动化运维。2.0网络架构中业务对网络QoS需求相比1.0网络架构提升一个数量级，接入带宽：25Gbps ~100Gbps，网络时延：50us~1ms，丢包率：0.1%~1%。

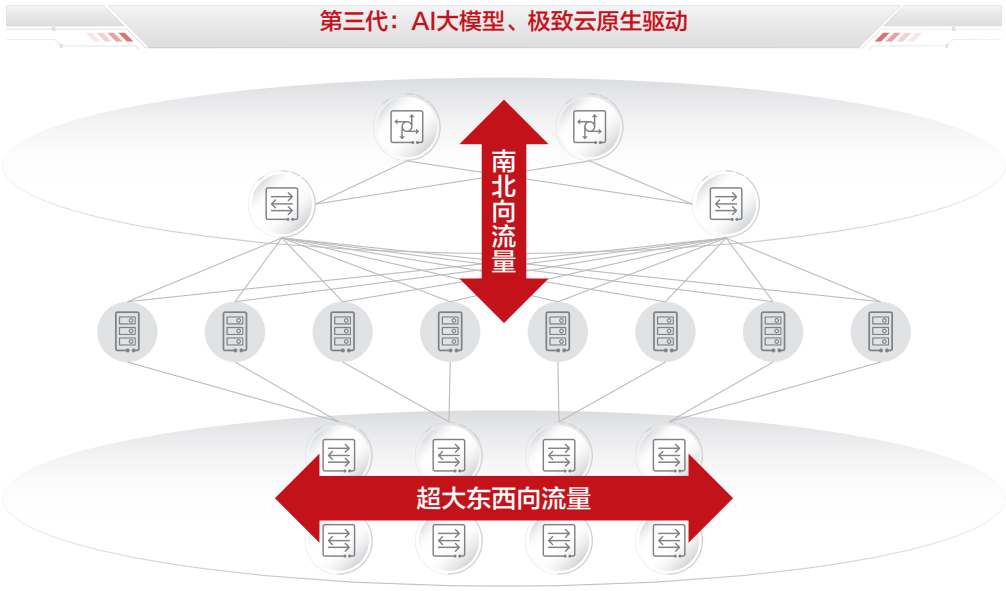


图15 数据中心网络架构3.0

数据中心网络架构3.0时代：相比2.0时代，东西向流量进一步加大，变为超大东西向流量，网络规模也变为超大网络规模。随着AI大模型业务的广泛部署，要求网络支持大带宽、高负载、无损网络，部署方式大规模、扁平化。并基于云原生技术栈，将网络的设计、部署、运维完全融入云原生生态，实现网络能力的极致弹性、自动化与分布式系统，即形成极致云原生驱动网络架构。3.0网络架构中业务对网络QoS需求相比2.0网络架构又提升一个数量级，接入带宽：>800Gbps，网络时延：10us~40us，丢包率：0。

数据中心网络架构持续演进，但存在如下问题：

- 采用分层网络架构（Underlay网络负责基础路由，Overlay网络负责逻辑网络），跨区域通信需经过多跳网关，导致延迟高（百毫秒级）、带宽不足（达不到Tb级转发能力）。同时，O/U网络间的解耦，让物理网络不能被充分利用，变相提升了通信成本。
- 租户缺乏数据中心网络物理拓扑和背景负载情况，在租户视角下的网络优化无法实现性能最优，多租户间可能存在网络使用冲突。
- S/O/U融合（Service/Overlay网络/Underlay网络深度融合）技术是华为云提出的数据中心网络架构创新方案，旨在解决传统云网络在规模、转发性能和智能化方面的瓶颈问题。

2) AI原生云广域网络发展趋势和需求

“东数西算”是国家战略工程，旨在通过构建全国一体化的算力网络，将东部密集的数据需求与西部丰富的能源结合，实现算力资源的优化配置。其核心依赖高性能、低延迟、智能化的广域网络支撑，其需求可归纳为以下维度：

- **超低延迟互联需求：**确保西部数据中心与东部用户间的数据传输延迟可控；
- **超大带宽需求：**支撑东西部间海量数据（AI训练集）的高效传输；
- **算力-网络协同需求：**实现“算力资源”与“网络资源”的智能匹配与动态调度；

针对上述需求，传统TCP/IP网络无法实现，存在如下的问题：

- **IP路由层：**完全基于邻居发现式路由通告&固定权重选路，缺乏局部路由拥塞的及时感知与恢复能力；
- **TCP传输层：**TCP流控窗口拥塞后恢复慢、收发两端TCP内核态升级困难，存在队头阻塞、TCP连接建立慢、网络迁移需要重建TCP连接等问题；
- **路由层/传输层/应用层之间的协同机制：**路由层无可靠QoS保障前提下，只能依赖传输层纠错和重传，加重了QoS体验劣化；同时由应用层进行Jitter消除、自适应速率编解码等处理；

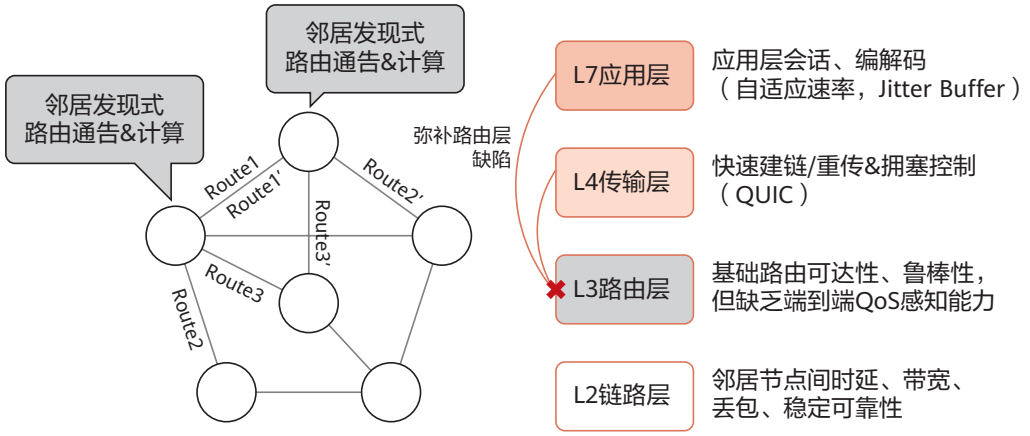


图16 传统网络存在的问题

针对上述问题，业界做了各种努力和尝试，但不能解决根本问题：

- **网络层MPLS-TE协议：**引入了端到端流量工程规划，但仍面向固定带宽分配，无法感知应用流量的带宽及QoS（丢包、时延）的动态变化。
- **网络层SRv6协议：**引入了端到端转发路径&行为可编程灵活性，但要求沿途段路由节点需要支持SRv6功能，全球云互联及跨异构云互联场景下难以保障该要求。
- **传输层QUIC协议：**解决了队头阻塞问题，单调递增的PN标识无需像TCP一样有序确认。优化了连接管理机制，初始建链过程更快，支持跨不同接入技术比如跨5G/WiFi的业务流连续性。QUIC协议改进了TCP的拥塞控制机制，提高了重发超时阈值准确度，通过增加冗余纠错码，降低了超时及丢包重传概率，并支持应用更敏捷的拥塞算法迭代。但QUIC协议涉及TCP应用生态的适配修改，存在与其他拥塞控制协议兼容的问题。

为了能让广域网络传输满足“东数西算”场景的需求，SDN网络架构是一个必然选择。Google是业界的先行者，Google推出的B4是业界部署的第一个数据中心互联SDN网络，采用Google自研交换机设备，运行纯IP网络，全球部署site数目55+。B4 SDN网络架构相比传统TCP/IP网络的优势：

- **保障高优先级业务QOS：**将应用的优先级纳入路由选路策略中，区分出高优先级和低优先级流量，调度保证高优先级流量低时延到达；
- **网络带宽利用率大幅提升（从平均30%~40%到近100%）：**使用非最短路径的包转发机制，喷洒式多路径转发，低优先级流量把空余流量挤满。

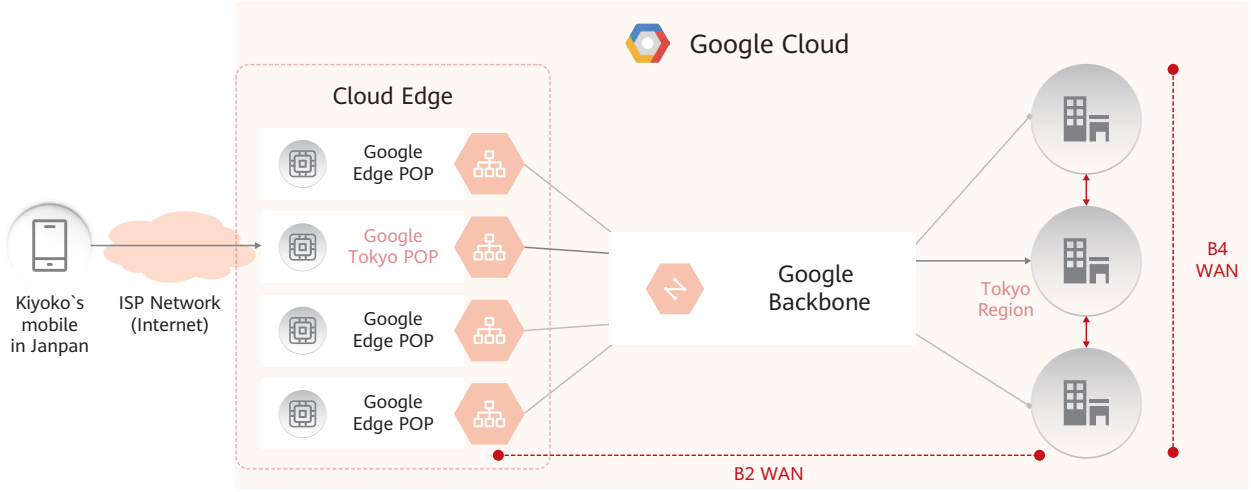


图17 谷歌数据中心互联SDN网络

» 关键特征与相关技术

1) AI原生云数据中心网络架构关键特征与相关技术

i. 无阻塞网络，满足高吞吐、零丢包需求

为了实现无阻塞网络架构，业界常用的技术包括：

- **Spine-Leaf架构：**用于构建数据中心网络，可支持流量无阻塞传输；
- **InfiniBand：**一种高性能计算和数据中心网络技术通信技术协议，实现GPU间高速通信；
- **RoCEv2：**一种基于以太网的远程直接内存访问（RDMA）技术，在以太网上实现高性能的数据传输和通信；
- **SRv6网络切片：**结合动态路由算法，依据网络实时状态和流量负载，智能选择最优传输路径。

ii. 确定性低时延网络，满足低时延、高可靠需求

为了实现确定性低时延网络，业界常用技术包括：

- **TSN（时间敏感网络）：**①IEEE 802.1Qbv等标准定义时间感知调度（TAS），为关键流量预留固定时隙；②优先级隔离：高优先级流量（如控制指令）可抢占低优先级流量（如文件传输）；
- **极低抖动：**①同步时钟：通过IEEE 1588（PTP）实现纳秒级时间同步，消除队列积累抖动；②流量整形：限制突发流量（如令牌桶算法），平滑发送速率。
- **高可靠性，实现网络可用性≥99.9999%：**①多路径冗余：快速故障切换；②确定性重传：基于时间窗的重传机制（如TSN 802.1CB）。

iii. 超节点网络

超节点网络是指将多个数据中心节点通过高速网络连接起来，形成一个逻辑上统一的资源池，该网络架构具有如下特点：

- **分布式架构：**物理上分布式部署，逻辑上统一，支持AI大规模分布式训练；
- **资源共享：**将超节点网络的计算资源(CPU/GPU/TPU集群)组成统一资源池，计算、存储和网络资源可跨节点共享与调配；
- **高带宽互联：**节点间通过超低延迟、高带宽网络连接，利用高速互联实现梯度同步和模型参数更新。

超节点网络使用的技术包括：

- **网络优化：**RDMA(远程直接内存访问)技术减少数据移动开销；网络计算软硬协同加速通信操作；自适应路由避免网络拥塞；
- **存储加速：**内存分级存储架构；计算存储融合设计；智能缓存预取策略；
- **软件栈支持：**分布式训练框架(如Horovod)的深度优化；超节点感知的调度器；拓扑感知的通信库。

超节点网络架构为AI业务带来的优势：

- **弹性扩展：**自动扩展/收缩训练或推理集群规模；按需增加节点，实现近乎无限的扩展能力，可支持千亿/万亿参数超大模型的训练；
- **高可用性：**单点故障不影响整体服务；
- **低延迟：**通过智能路由和边缘计算降低延迟；
- **负载均衡：**根据AI工作负载需求动态分配计算资源，并支持突发性AI工作负载需求。

2) AI原生云广域网络架构关键特征与相关技术

针对传统广域TCP/IP网络的不足，华为云推出了广域网络架构：应用传送网络ADN（Application Delivery Network）。

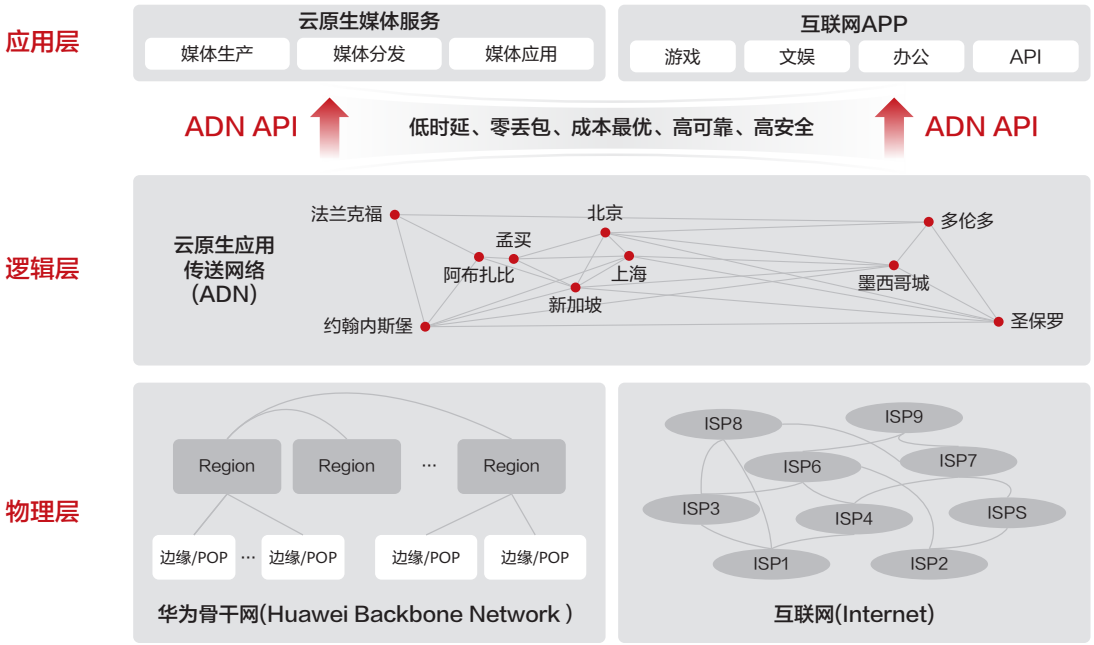


图18 应用传送网络ADN

如图18所示，ADN为云原生服务，乃至更广义的云上互联网应用提供了多级QoS，高可靠，高弹性的网络基石，相比基于“尽力而为”的使用IP路由转发的Internet互联网，ADN网络是一张叠加在Internet互联网，以及华为云遍及全球的云端及分布式边缘基础设施和专线网络之上的Overlay网络。

ADN网络彻底解决了互联网缺乏QoS保障，局部路由拥塞收敛慢，以及专线成本高，覆盖区域受限的问题，具备软件定义的可编程能力，无需升级改动存量运营商网络，即可支持分钟级新增路由节点及路由变更，使得网络具备了云的“弹性敏捷”的核心特征，从而为业务提供了兼具互联网全域覆盖、低成本及专线的确定性QoS保障优势的基础网络传送服务，并且可支持应用驱动的SLA与QoS。

i. ADN网络的三大核心技术特征

- **广覆盖、高敏捷、全互联的网络拓扑：**通过无所不在、彼此互联互通，超过2000个ADN节点的全球广泛覆盖，ADN网络实现了最终用户的一跳入网；同时，通过支持ADN节点的分布式容器化部署，实现了分钟级节点增删与网络拓扑更新的高弹性、高敏捷；通过任意ADN节点之间基于Full Mesh的点到点测量，为任意2个ADN节点之间动态最优路径的选择提供了依据和保障。
- **多目标驱动智能路由，多样化接入协议传输：**支持千节点分钟级端到端路由图优化算法，实现智能路由计算；支持单流分多流、多流合并单流，具备多优先级路径的实时选择能力；具备抗弱网协议增强、具备高可靠的传输能力，实现智能拥塞控制；通过华为自研设计的nStack协议栈，DPDK／用户态驱动转发的技术，实现近线速的Overlay转发能力；提供了TCP/UDP/域名解析/SDK模式等灵活多样化的ADN网络接入协议选项。

- **应用驱动、软件定义的SLA，租户和业务感知的流量调度：**在ADN的API定义中，通过应用驱动、软件定义的网络层/传输层/应用层QoS/SLA指标，比如网络层的时延、丢包，以及媒体应用层的抖动、音视频MOS等，描述上层应用App希望ADN网络达成的质量保障水平及目标；在应用感知方面，基于云服务类型感知的网络流量预测，及基于AI、大数据统计的租户应用流量画像，ADN网络进一步支持业务流量的分时错峰调度，以及跨端边云的应用与数据迁移同步能力。

ii. ADN网络给AI业务带来的优势

- AI业务的云租户可从各运营商的城域网经由分布式单线IP就近接入到分布式边缘站点，再通过分布式边缘站点经由物理专线连接到主Region服务区的云服务、云主机、云容器实例，由于动态BGP与单线IP在国内定价差价达近10倍，使得公有云的网络接入总成本降低达40%以上。
- ADN网络是应用驱动的网络，可为AI应用提供端到端访问QoS保障，实现云端、云边访问零丢包、降时延超过25%；在跨海场景下，业务加速效果尤为明显。
- “东数西算”场景，通常情况下AI数据采用的是CC云连接传输，云连接租户独占固定带宽，存在着闲时浪费的情况。采用ADN网络传输后，实现多租共享弹性带宽，消峰错谷，可以降低传输成本超过40%。

3.2.5 华为云AI-Native云基础设施实践

» 3.2.5.1 CloudMatrix多元算力超节点HPS

据咨询公司的预测，到2030年，全球每年产生的数据总量将达到1YB，相比2020增长23倍，其中通用算力将增长10倍，AI算力增长500倍。AI大模型、AIGC、媒体渲染、大数据和数仓平台、基于云的仿真和超算等广泛存在的多样性、紧耦合、大规模应用的计算范式。

单一类型的计算资源，单一节点的计算能力、存储能力，以及配比固定、松散协同的扩展模式已经难以满足日益复杂且快速变化的应用部署需求。新型应用驱动计算范式从单算力向混合算力协同发展，从单机向集群灵活部署发展，从传统应用松散分布向多样应用紧密融合发展，对未来数据中心架构提出新的诉求。

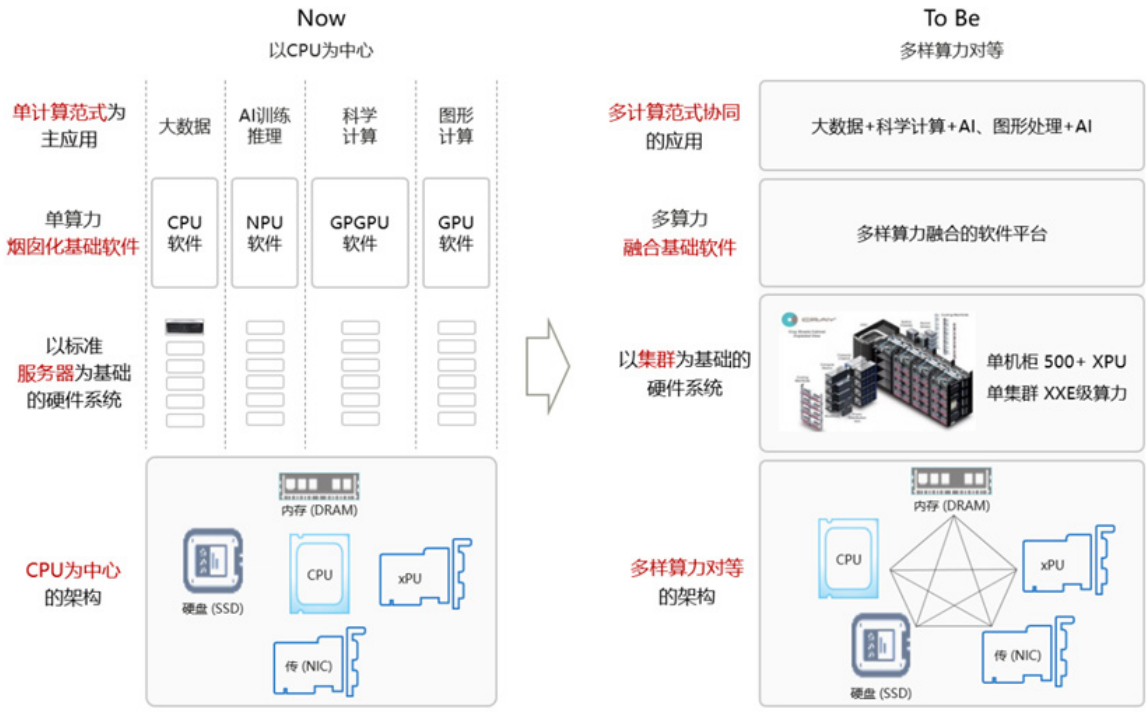


图19 集群架构变化

1) 高效组合异构算力，提升任务处理速度：

在服务器、机架和集群内，需要高效组合不同算力类型完成计算任务，打破传统主从式结构，实现设备之间直接的互联互通，使计算任务执行更快，资源利用率更高

架构需要具备如下特征：

- 异构算力以对等方式横向扩展，各组件之间可直接通信，互相调用。
- 通信带宽高、时延低，支持在细颗粒度任务上做并行处理或调用。
- 总线机制支持多颗芯片协同完成单个功能调用。

2) 资源动态分配，提升资源利用效率

提高数据中心资源利用效率，需要打破服务器盒子边界，实现大范围不同设备的池化和资源动态分配。改变以单节点能力为上限的资源分配方式，总线设备和内存资源不固定归属于特定计算单元，在总线层面实现资源动态注册和分配，满足资源高效应用需求。

3) 快速异常恢复，提升系统可靠性和可用性

- 基于高速互联总线，实现内存紧急借用，业务上下文秒级迁移，实现OS宕机场景核心业务中断下降90%。
- 提供极速热迁移能力，实现业务无感(<50ms)完成OS冷补丁和硬件故障维修。
- 池化资源故障秒级快速检测，跨层故障秒级快速通知，层次化细粒度隔离(页隔离、节点级隔离)。
- 控制设备单点故障均可冗余秒级切换恢复业务，相对于传统单网卡接入和单控制面，均从单点切换到高可靠架构。
- DPU通过冗余DB卡，实现单卡故障秒级切换，训练推理业务不中断。

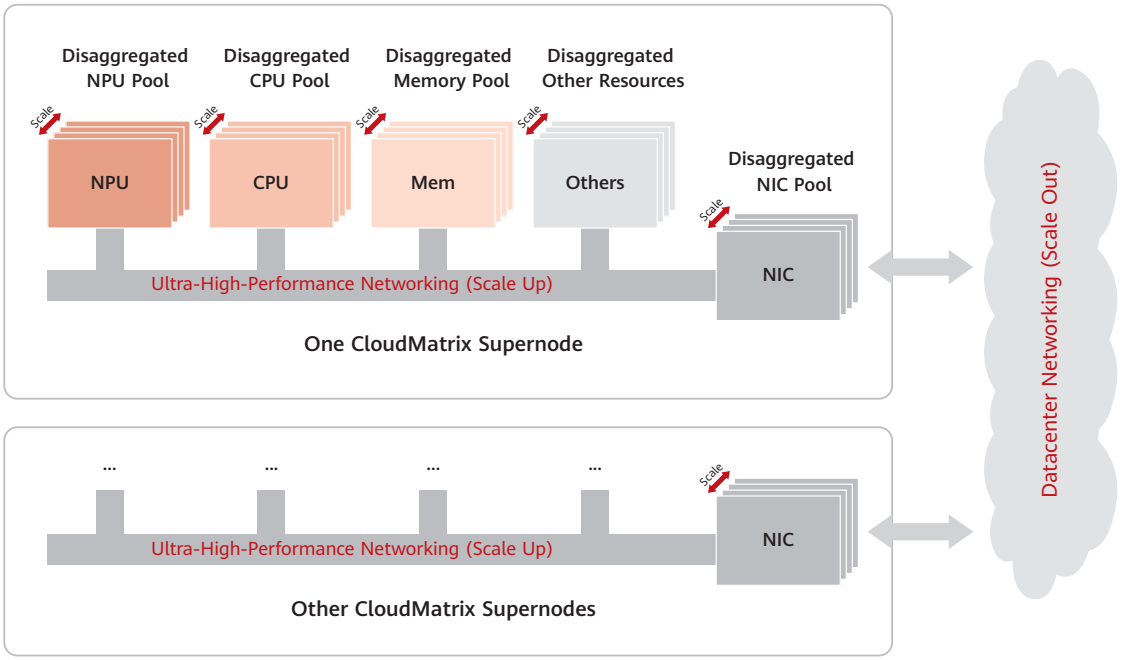


图20 CloudMatrix 架构示意图

» 3.2.5.2 华为CloudMatrix超节点

如上图概述，CloudMatrix超越了传统的以CPU为中心的层次化设计，促进所有异构系统组件（包括NPUs、CPUs、DRAM、SSDs、NICs以及专用加速器）之间的直接高性能通信，而无需CPU中介。该架构的核心是超高带宽、低延迟的统一总线（UB）网络，它促进了系统范围内的高效数据移动和协调。在此互联基座之上，CloudMatrix提供了四个基础能力，共同定义了AI原生基础设施的新范式：

- 可扩展的TP/EP通信：UB互联支持NPUs之间的直接高吞吐点对点通信，使TP和EP组能够超越单个节点的限制。这消除了节点间瓶颈，允许大模型在supernode上高效分布；

- 异构工作负载的灵活资源组合：CloudMatrix将CPU、NPU和内存分离为独立的池化资源，实现基于工作负载需求的细粒度、工作负载驱动的组合。这种灵活性允许根据工作负载需求以细粒度分配资源，例如内存丰富的缓存节点、CPU密集型预处理节点，使部署摆脱固定的节点配置或基于PCIe的主机设备耦合；
- 融合工作负载的统一基础设施：UB网络的高带宽支持在同一规模扩展的基础设施中同时运行AI和数据密集型应用。这使得推理、训练、仿真和分析等LLM工作负载能够融合执行；
- 通过分离式内存池实现内存级存储性能：CloudMatrix将集群中CPU附加的DRAM聚合为一个共享的高性能内存池，通过UB访问。这一基座支持弹性内存服务（EMS）等，通过消除传统的I/O瓶颈来加速KVCache重用、参数加载和模型检查点等延迟关键操作。

CloudMatrix384的一个关键特征是其完全点对点、完全互联的超高带宽网络，通过UB协议将所有NPUs和CPUs连接起来，如下图所示。CloudMatrix384的UB设计是中提出的UB-Mesh的前身。每个Ascend芯片通过UB交换机连接，实现节点间通信性能接近节点内水平。节点间带宽退化小于3%，节点间延迟增加不到1微秒。由于现代AI工作负载主要是带宽密集型而非延迟敏感型，这种微小的延迟开销对AI任务的端到端性能影响可以忽略不计。总体而言，这种设计使CloudMatrix384能够作为一个紧密耦合的大型逻辑节点运行，具有全局可寻址的计算和内存，促进统一资源池化和高效工作负载编排。

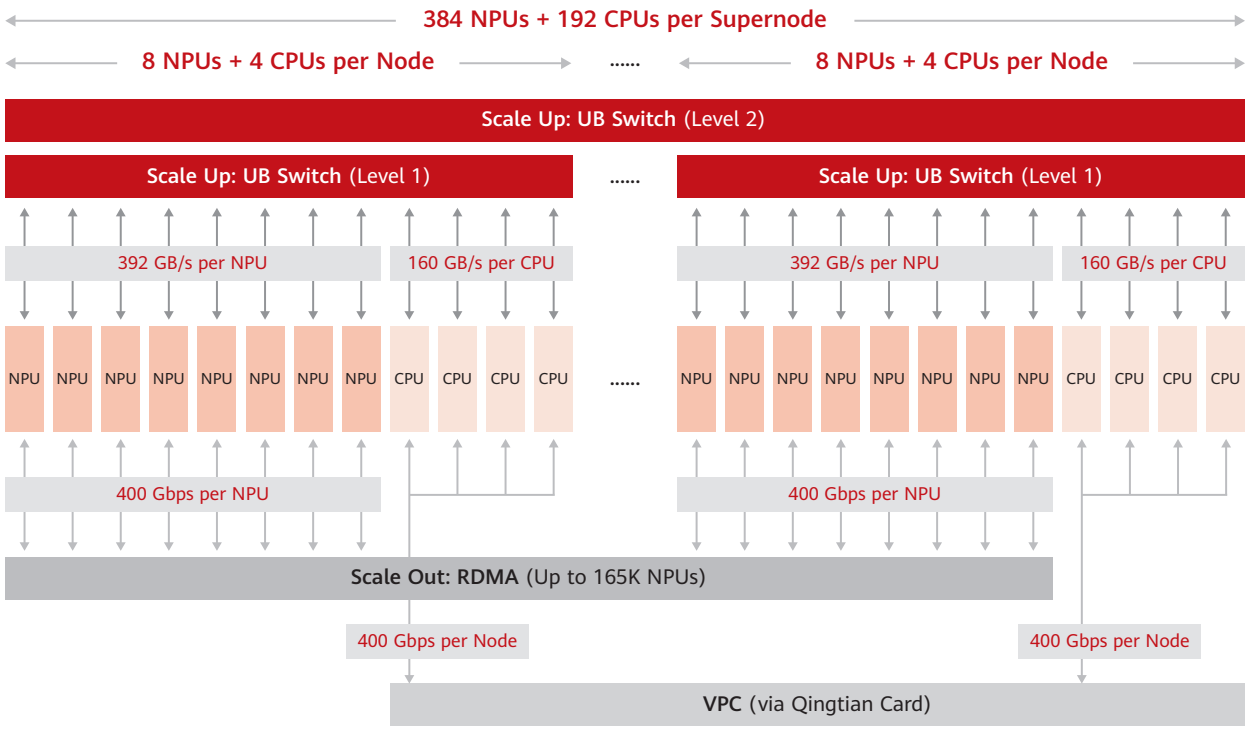


图21 点对点硬件架构

为了支持多样的流量模式并保持与传统数据中心网络的兼容性，云矩阵384融合了三个不同但互补的网络plane：

- UB Plane：UB Plane是超级节点内主要的超高带宽扩展结构。它以无阻塞的All-to-All拓扑直接连接所有384个NPUs和192个CPUs。每个Ascend 910C芯片贡献超过392 GB/s的单向带宽。UB plane支持：（1）高效实现细粒度并行策略，如TP（张量并行）和EP（专家并行），不受节点边界的限制；（2）快速的点对点访问池化内存（涵盖CPU和NPU内存），这对于高效缓存模型权重和KVCache至关重要。

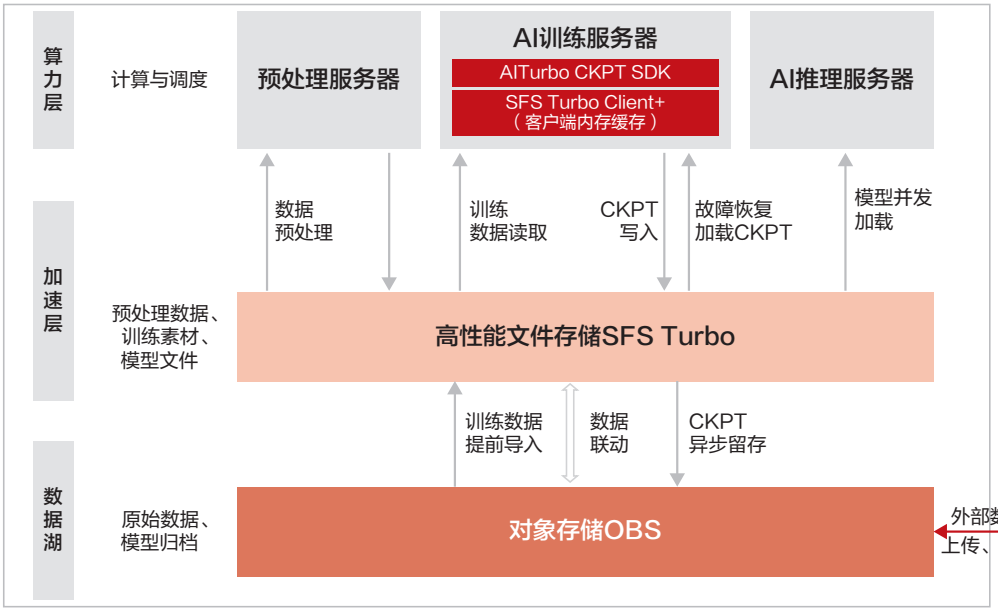


图22 华为云AI-Native智算存储训练加速解决方案

- RDMA Plane：RDMA Plane实现云矩阵384超级节点与外部RDMA兼容系统之间的扩展通信。目前采用以太网收敛的RDMA（RoCE）以确保与标准RDMA堆栈的兼容性。每个NPU提供高达400 Gbps的单向RDMA带宽。只有NPUs参与此plane，将RDMA流量与控制和存储操作隔离。主要功能包括：（1）推理过程中Prefill和decodeNPU间活跃KVCache数据的高速传输；（2）支持使用RDMA兼容框架的分布式训练和推理；（3）多集群部署中超级节点间的低延迟互连。
- VPC Plane：VPC plane通过高速网卡（擎天）将CloudMatrix连接到更广泛的数据中心网络，每个节点提供高达400 Gbps的单向带宽。它基于标准的以太网和IP协议运行，可选支持UB over Ethernet（UBoE）。VPC plane处理：（1）管理操作和控制plane操作，如部署、监控和调度；（2）对持久存储的访问，包括对象存储服务（OBS）、弹性卷服务（EVS）和可扩展文件系统服务（SFS）；（3）CPU驻留工作负载的外部服务通信，例如数据库和用户界面。

» 3.2.5.3 EMS&SFS Turbo AI原生云存储

华为云面向AI场景推出了AI-Native智算存储解决方案，提供基于对象存储服务OBS+高性能文件服务SFS Turbo+弹性内存服务EMS（Elastic Memory Service）的AI-Native智算存储加速方案。

1) AI 训练加速方案

- 以对象存储OBS数据湖作为统一数据底座，对象存储OBS提供HDFS/S3/POSIX多协议访问同一份数据，和数据接入、大数据处理、内容审核等高阶服务无缝集成，高效衔接AI系统各个工作环节，避免数据在各工作环节之间进行拷贝搬迁，避免数据冗余存储多份。同时，对象存储OBS提供标准/低频/归档/深度归档等多种存储类别，结合数据生命周期管理，解决AI场景中海量数据长期高可靠低成本存储。

- 在训练推理等对存储性能要求极高的环节，为了更好的加速大模型训练和推理，提供高性能文件服务SFS Turbo加速层存储作为OBS数据湖存储的补充。SFS Turbo可以提供亚毫秒级的数据访问延迟、千万级的IOPS和TBps级别的吞吐能力，有效提升数据清洗、大模型训练、及推理中模型加载的效率。SFS Turbo高性能文件和OBS数据湖之间集成了数据联动功能，无需借助外部工具，即可实现数据高效流转。同时，SFS Turbo推出了三级缓存加速架构，该架构基于SFS Turbo高性能文件存储服务端，SFS Turbo Client+内存缓存客户端，及专门针对AI场景中CheckPoint任务快照保存与恢复等AI语义进行加速的AITurbo SDK技术组件，为大模型训练构建了高效的存储方案，通过对应用层AI生态的理解和端到端全栈优化，实现AI场景千模百态的全面加速。

- 华为云推出了全球首创的弹性内存服务EMS（Elastic Memory Service），一种以DRAM内存为主要存储介质的云基础设施服务。通过EMS，华为云将传统的“计算-存储”分离的两层云架构升级为“计算-内存-存储”分离的三层云架构，其中新增的“内存层”即为EMS。这种新型的三层云架构能有效解决存力痛点，从而具有高资源弹性、高资源利用率和高性能等优势。以下介绍关键技术及价值：

  - 数据联动技术：SFS Turbo高性能文件存储内置Bucket Link数据联动功能，SFS Turbo里的文件系统可以绑定容量层的OBS对象桶，用户无需手工部署外部迁移工具即可实现在OBS对象存储和SFS Turbo高性能文件存储两个分布式存储服务之间进行高速数据流动，存储各节点均参与数据导入、导出，数据流转比人工带外部署迁移工具方式更加简洁高效。大模型训练过程中周期性产生的CheckPoint数据可以高速写入SFS Turbo高性能文件缓存，减少对上层训练任务的中断和阻塞，可以提高CheckPoint保存频率，减少训练任务故障时需要从最近一次CheckPoint重新训练的损失。同时，SFS Turbo高性能文件缓存自动以异步方式将CheckPoint导出到关联的OBS对象存储桶中进行长期低成本存储。最后，SFS Turbo高性能文件存储通过配置缓存数据淘汰功能，及时将长期未访问的数据从缓存中淘汰，释放SFS Turbo高性能缓存空间。

  - 三级缓存加速技术：SFS Turbo高性能文件缓存加速层存储提供L1服务端内存缓存，L2客户端内存缓存，及针对CheckPoint保存与恢复等场景进行加速的L3 AITurbo SDK，形成三级缓存加速技术，加速AI训练过程中的训练数据集读取，CheckPoint快速保存及故障时的CheckPoint快速恢复。

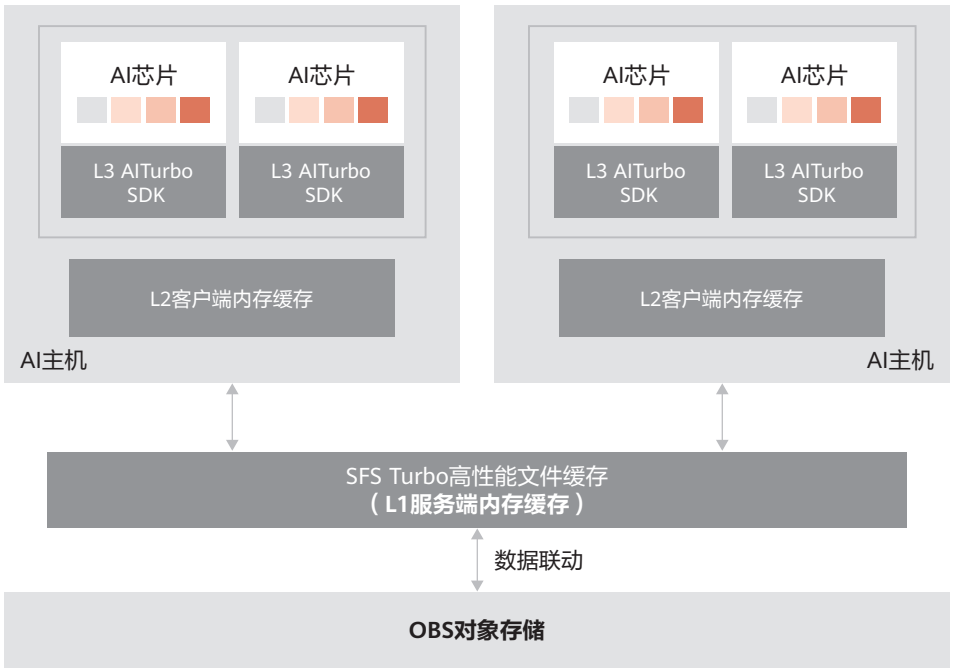


图23 三级缓存加速技术

训练数据集访问加速：在业务访问数据集文件时，SFS Turbo高性能文件存储会将NVMe SSD存储池中的数据文件缓存到L1服务端分布式内存缓存中，减小AI训练访问数据集的时延，同时在大规模训练集群并发访问数据集时，以充分发挥L1服务端内存缓存带宽优势，实现比NVMe SSD硬盘层更大的吞吐能力。另外SFS Turbo高性能文件存储的分布式元数据，可以支撑百亿级小文件扩展，进一步缩短了海量小文件元数据操作的时延，提升了海量小文件操作的IOPS吞吐。

CheckPoint保存及恢复加速：SFS Turbo提供的L3 AITurbo CKPT读写加速组件针对进程级故障和JOB任务级故障等场景，对接PyTorch/MindSpore/DeepSpeed等主流大语言模型训练框架，专门针对AI训练中的CheckPoint保存及恢复过程进行加速，实现 CheckPoint先高速同步写到本机L2客户端内存缓存，再异步持久化到服务端存储，最大程度减少CheckPoint同步保存耗时，减少了训练任务中断阻塞。AI训练任务发生进级故障时，利用本机SFS Turbo Client+的L2客户端内存缓存实现CheckPoint原地秒级快恢，发生节点故障及JOB任务重调度场景下，利用客户端节点间高速参数面网络实现CheckPoint广播技术加速CheckPoint恢复速度，最大程度减少CheckPoint并发恢复耗时，避免训练任务故障恢复时由于远端存储带宽瓶颈导致长期阻塞。SFS Turbo通过 L3 AITurbo CKPT读写加速组件及L2客户端内存缓存功能，可以有效加速CheckPoint保存及恢复速度，可以提高CheckPoint保存频率，大幅减少故障恢复时需要从上一次CheckPoint重新训练的损失，同时CheckPoint保存和恢复加速减少了大规模AI集群算力的空闲损失，提高了AI集群可用度，加速了AI训练任务进程，确保大模型训练能够按时完成，节省出的算力可以训练出更多更新的大模型。

2) AI 推理加速方案

为了解决云基础设施中存在的存力痛点，华为云推出了全球首创的弹性内存服务EMS（Elastic Memory Service），一种以DRAM内存为主要存储介质的云基础设施服务。华为云将传统的“计算-存储”分离的两层架构升级为“计算-内存-存储”分离的三层架构，这种包含EMS的新型三层云基础设施架构具有高资源弹性、高资源利用率和高性能等优势，能够有效解决前述三大存力痛点：

- 针对AI场景中“持久化存储性能不足”的问题，EMS作为计算层与存储层之间的高性能缓存层，利用DRAM介质缓存来自HDD和SSD介质的数据，显著提升数据访问速度。
- 针对AI场景中“DRAM利用率低”的问题，EMS将AI服务器中的DRAM资源进行解耦并池化，形成EMS内存池。EMS内存池中的DRAM资源根据不同计算任务的需求进行动态分配，从而实现内存资源的高效利用。计算层与内存层之间通过华为专有的高性能网络总线连接，确保内存资源解耦池化后的高访问性能。
- 针对AI加速器中的“显存内存墙”问题，EMS利用内存池中的DRAM资源扩展AI加速器的显存内存，通过增加DRAM容量来扩展显存容量，并利用DRAM带宽补充显存带宽，从而大幅提升AI训练和推理的整体性能。

下面将首先阐述EMS的软件架构，随后探讨EMS内存解耦池化的关键技术，最后介绍EMS针对不同AI场景的内存加速关键技术。EMS的软件架构主要由三部分组成：领域专用服务SDK、分布式内存池和管理控制面，如图所示。EMS的软件面向高易用性、高弹性扩展性和高可用性设计。

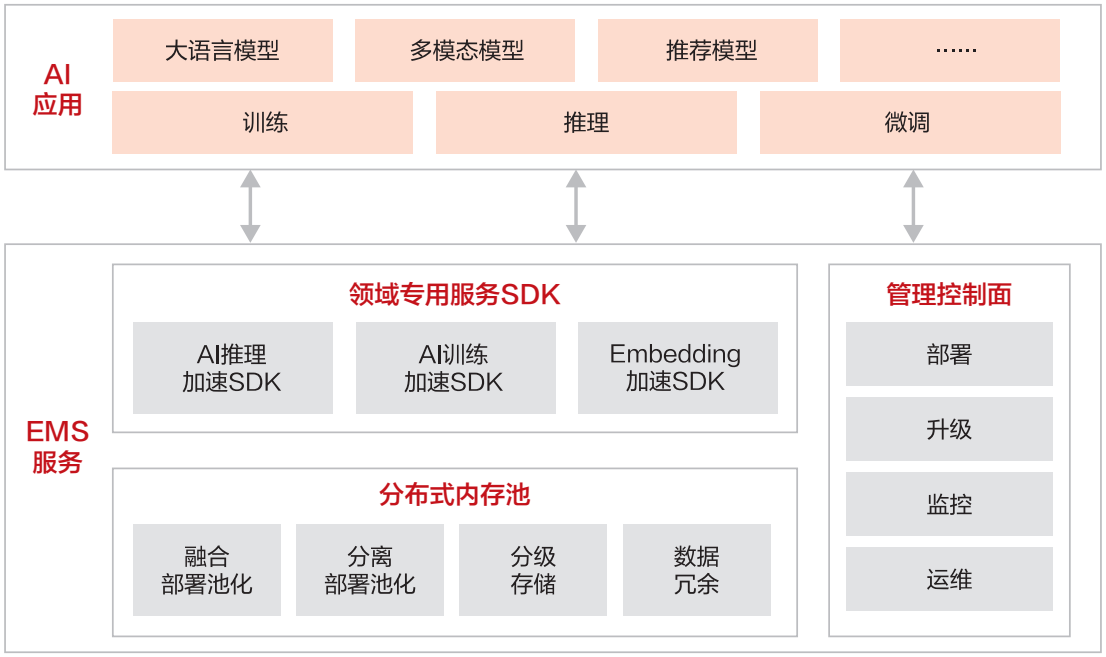


图24 EMS软件架构

领域专用服务SDK包含一系列面向不同AI应用场景的插件和接口服务SDK，提供业务系统接入、业务数据布局和近数据处理等功能，实现业务请求的内存加速。目前，该技术主要应用于大语言模型、多模态模型、推荐模型等的训练和推理，通过分布式内存池提升处理效率并降低成本。

分布式内存池负责跨节点的内存空间管理、数据负载均衡和数据恢复等任务，通过空间池化、共享访问和故障切换等机制，确保系统具有低成本、高扩展性和高可用性。内存池提供两种部署模式：（1）融合部署，即利用AI服务器中的DRAM，将DRAM内存池化以实现分布式共享，并进行本地亲和的调度和访问；（2）分离式部署，即使用独立内存服务器提供内存池空间，通过高速内存总线实现对内存池空间的访问。

管理控制面负责服务的部署、监控、升级及运维管理等功能，通过华为云的云原生基础设施为用户提供一站式的云上运维解决方案。

下面介绍分布式内存池及领域专用服务SDK的具体技术。

i. 内存解耦池化

在AI训练和推理场景中，AI服务器的数量可达数千至数万台，每个服务器上的DRAM内存通常按照最大需求进行配置和预留。然而，实际操作中，服务器间的内存利用率往往不均衡。由于DRAM是AI服务器成本的重要组成部分，池化DRAM以提高利用率和降低成本显得尤为重要。此外，随着“显存内存墙”问题的日益凸显，EMS通过在显存和DRAM之间进行卸载及相应的数据管理来解决这一问题。EMS内存池需要满足池化和卸载两个关键需求，具体技术将在后续章节详细介绍。

- **内存池融合部署架构**

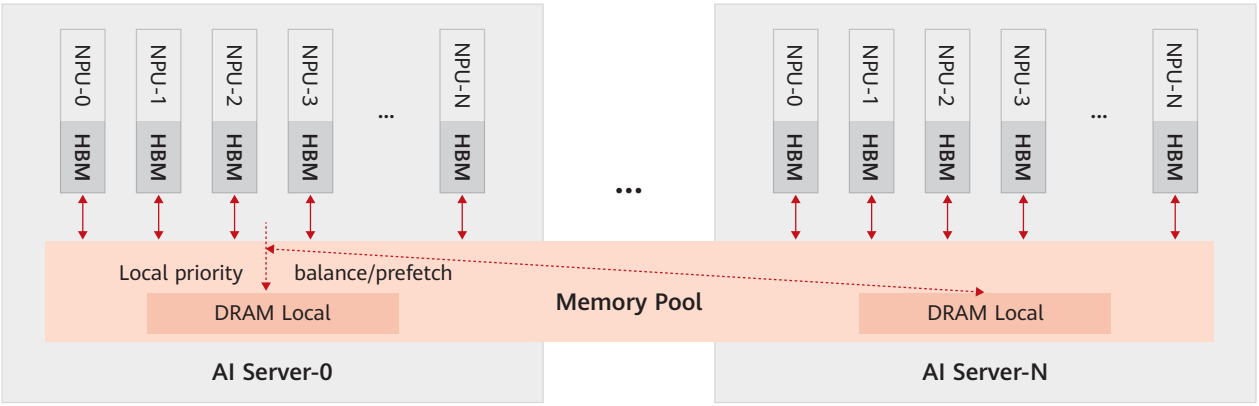


图25 内存池融合部署架构

内存池融合部署架构通过利用AI服务器内的本地DRAM进行池化管理，提升内存利用率，如图25所示。内存池化还带来了共享能力，使得卡间能够进行高效的数据共享，具体技术点如下：

1) 服务器内DRAM池化：AI加速卡根据需求从内存池中分配DRAM，避免了按卡粒度预留导致的卡间使用不均和利用率低的问题。

2) 服务器间DRAM池化：整个AI服务器集群形成一个大的内存池，解决了服务器间内存利用率不均的问题，提高了利用率并降低了成本。此外，任意AI节点能够访问任意缓存数据的能力，为AI训练和推理场景下的加速技术（如基于内存CheckPoint的故障快速恢复、长文本推理和PD分离等）的应用提供了支持。

3) 数据访问亲和性调度：在典型场景下（如大模型训练CheckPoint和KV Cache存储），通过亲和感知、动态均衡和预加载的方式，确保训练和推理过程中的高带宽内存访问需求，并在一定程度上解决了内存利用不均衡的问题。

- **内存池分离部署架构**

基于高速网络总线加速的分离式内存池，是EMS内存池化和容量卸载的最终形态，如图26所示。通过引入DRAM专有服务器硬件最大化降低成本，在内存使用均衡性、利用率和共享方面都提供了最优解。对外以一个整体展示，即体现在整柜硬件交付，也体现在如全局地址空间的空间管理能力。

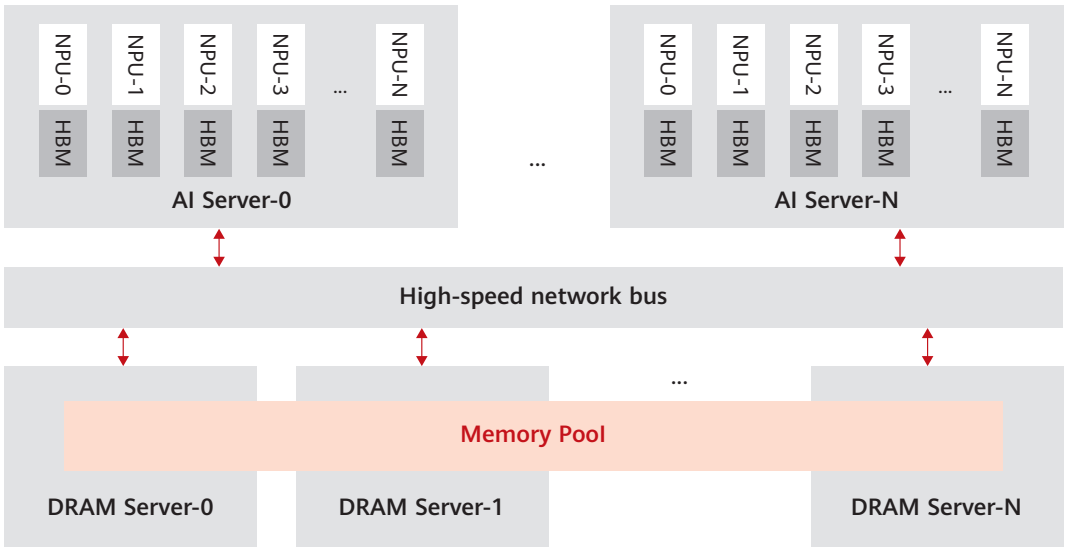


图26 内存池分离部署架构

- 分级存储

通过DRAM卸载解决显存的内存容量墙问题，但在某些场景下，DRAM同样面临容量不足和成本过高的问题。EMS进一步将数据卸载到高速持久化存储介质（如SSD），最终形成多级的分层卸载存储形式，如图27所示。通过引入AI训练和推理流程感知的算法，进行显存和DRAM之间、DRAM和SSD之间的主动卸载和取回调度，使得在训推效率和成本上达成平衡。

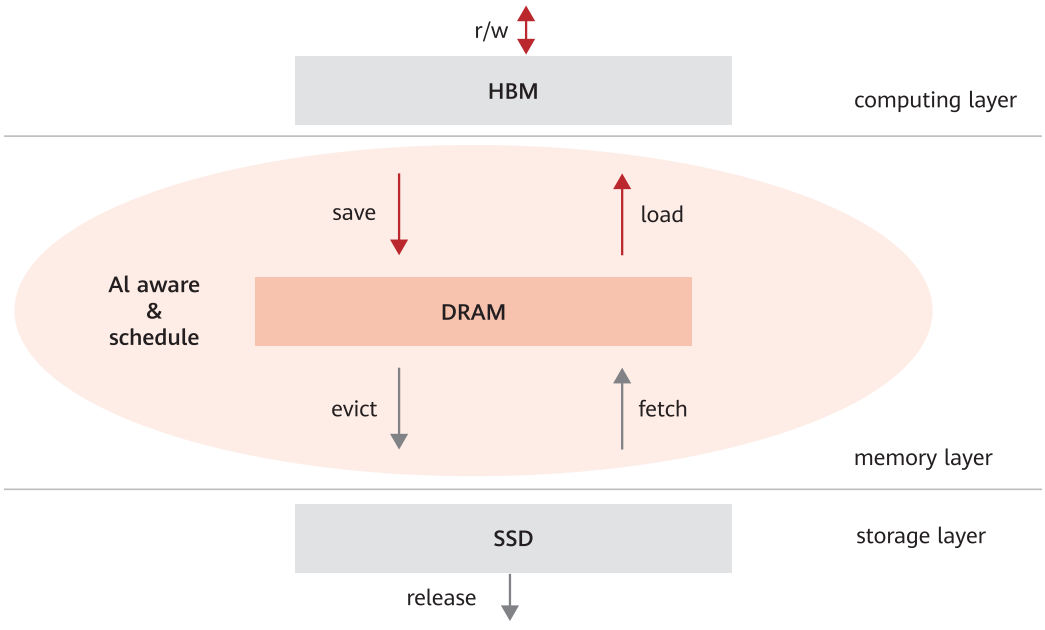


图27 分级存储技术示例

- 数据冗余

在云数据中心中，通常拥有成千上万的AI服务器。EMS为如此大规模的AI服务器提供弹性内存服务，也会具有较大的分布式规模。AI训练和推理过程中卸载到EMS中的数据如果丢失，将会造成AI任务的中断或重新执行。EMS内存池提供基于副本和纠删码的内存数据冗余能力，以大幅提升数据的可用性。

ii. 面向AI推理的加速技术

在前面章节已阐述，Transformer模型推理中存在严重的显存内存墙问题。为解决这一问题，EMS提供了以下三种技术以加速AI推理：以存带算、显存扩展和计算卸载。

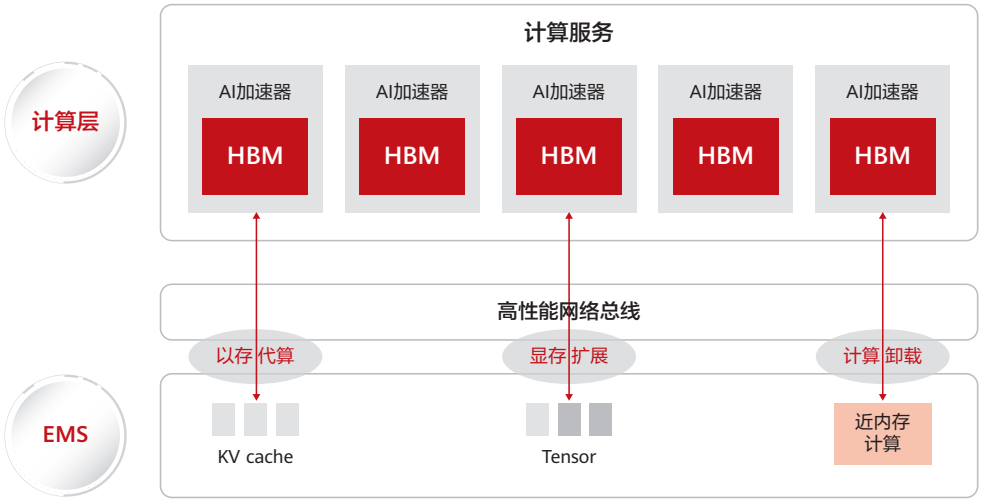


图28 AI推理场景中的EMS关键技术

- 以存代算

在Transformer模型的推理过程中，由于AI加速器的显存内存容量限制，现有的推理系统无法在AI加速器的显存中持续保存多轮对话的KV缓存。为了应对这一问题，系统通常会丢弃已处理对话的KV缓存，腾出显存空间来服务新的请求。然而，当这些被丢弃的KV缓存对应的对话再次出现时，系统必须重新计算这些KV缓存。这种重复计算不仅浪费了计算资源，还增加了推理成本。

为了减少成本并提升推理性能，EMS服务引入了以存代算技术CachedAttention。该技术利用EMS中的大容量多级内存池来存储和复用多轮对话中产生的KV缓存。具体操作是，当一个会话变为非活跃状态时，将相应的KV缓存保存到EMS中。当该对话重新被激活时，再从EMS中加载并复用这些KV缓存，从而避免了重复计算。此外，EMS还采用了以下技术来优化缓存系统性能：（1）采用逐层预加载和异步保存策略，以减少加载和保存KV缓存的时间；（2）利用多级缓存结构，通过更大容量的存储介质提供充足的缓存空间；（3）通过自动感知调度器中的任务队列信息，实现多层次存储介质间的缓存调度，以提高访问效率；（4）将位置编码从KV缓存中分离，确保在模型上下文窗口长度溢出时KV缓存的可重用性。

通过以存代算技术，EMS有效地避免了多轮对话中的重复计算，显著降低了首字时延，提高了预填充阶段的吞吐量，并降低了端到端的推理成本。

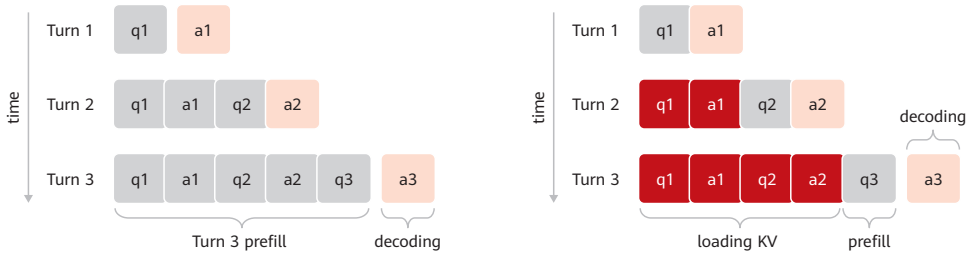


图29 多轮对话中使用EMS

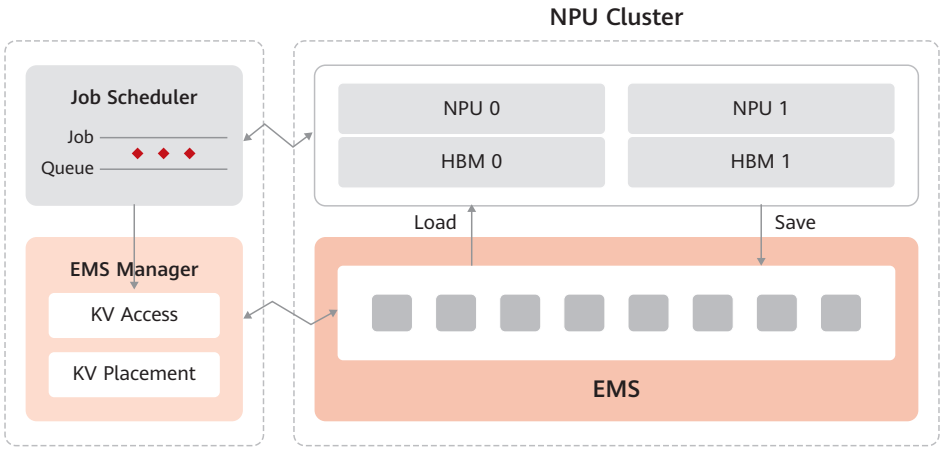


图30 EMS以存代算技术

- 显存扩展

由于AI加速器中的显存内存容量限制，AI加速器可能无法容纳大的模型，或者即使能够容纳，也无法使用较大的批处理大小（Batch Size）。为了解决这一问题，EMS采用了显存扩展技术，以增加AI加速器的可用显存，从而支持运行超出显存容量的模型或增加推理的批处理大小。在推理过程中，EMS将显存中的KV缓存、模型权重等数据动态卸载到一个大容量的共享弹性内存池中，如图31所示。通过利用计算层与内存池之间的高性能网络总线，EMS实现了数据传输与计算过程的流水线并行，有效减少了内存池的访问开销。得益于这种大容量、高性能、共享访问的弹性内存池，EMS的显存扩展技术能够增加推理的批处理大小，进而提升AI推理的整体吞吐率。

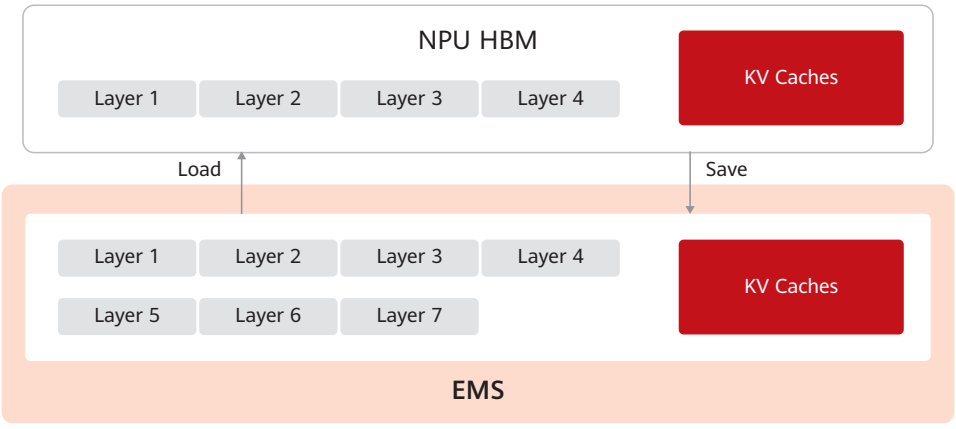


图31 EMS显存扩展技术

- 计算卸载

针对Transformer模型推理中遇到的显存内存墙问题，EMS通过计算卸载技术将自注意力模块相关的KV缓存数据和轻量级算子卸载到内存池中，利用EMS中的DRAM容量扩展显存的容量，并通过DRAM的内存带宽补充显存的带宽。具体而言，Transformer模型的自注意力模块需要加载整个推理上下文的KV缓存以完成注意力分数的计算，这一过程涉及大量KV数据的读取，而相关算子的计算量相对较小。与此相反，前馈网络模块主要由计算需求较大的全连接网络算子构成，对存储容量的需求较小。EMS根据这些不同的计算特性，将自注意力模块和前馈网络模块分别在计算能力较小但存储能力较大的CPU侧和计算能力较大但存储能力较小的AI加速器上完成。同时，EMS根据推理任务的服务级别协议（SLA）需求，智能地决定卸载的时机和粒度，通过弹性内存池的大容量和大带宽优势缓解了显存内存墙问题，采用异构推理方案提升了AI推理的端到端性能和性价比。

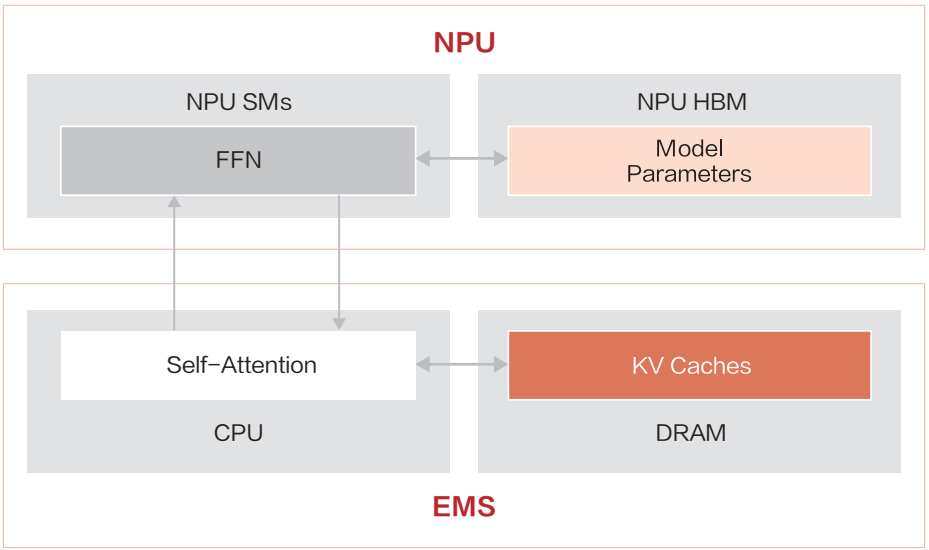


图32 EMS计算卸载技术

iii. 面向推荐模型的加速技术

推荐模型在广告、新闻和视频推荐等多个领域得到了广泛应用。与大型语言模型和视觉模型不同，推荐模型的输入特征中包含大量ID类型的数据，如用户ID、新闻ID和视频ID等。这些ID特征通常具有极高的维度且非常稀疏，难以直接被神经网络处理。为了解决这一问题，推荐模型采用了Embedding加多层感知器（MLP）的架构。通过Embedding技术，ID类型特征被映射到低维向量，从而缩小了与神经网络之间的差距。推荐模型中包含多个Embedding表，每个表负责将特定类型的ID特征转换为Embedding。推荐模型的前向计算过程：ID特征首先通过Embedding表转换为Embedding，然后这些Embedding经过池化处理，如求和或平均，最终输入MLP以生成最终的推荐标签。由于ID类型特征的基数庞大，推荐模型中Embedding层的参数量非常大。例如，亿级的视频ID特征可能需要一个拥有万亿参数的Embedding表。因此，Embedding通常占据了推荐模型中超过99.9%的参数。

与计算机视觉和大语言模型不同，推荐模型训练面临的一个主要挑战是数据更新迅速且特征极其稀疏，例如高维的one-hot或multi-hot向量。此外，新数据的不断加入可能导致特征向量达到百万甚至亿级，使得模型规模达到万亿字节，难以存储在单个NPU或单台机器的内存中。在训练和在线推理过程中，神经网络计算中涉及的Embedding数据量相对较小，通常不到总数据量的1%。因此，业界普遍采用一个独立的全局Embedding存储服务（或参数服务器）来提供统一的Embedding管理、存储和计算能力。EMS作为推荐模型训练和推理过程中的Embedding存储服务，旨在实现高资源利用率和高的访问性能。EMS提供Embedding存储服务的关键技术将在下文中介绍。

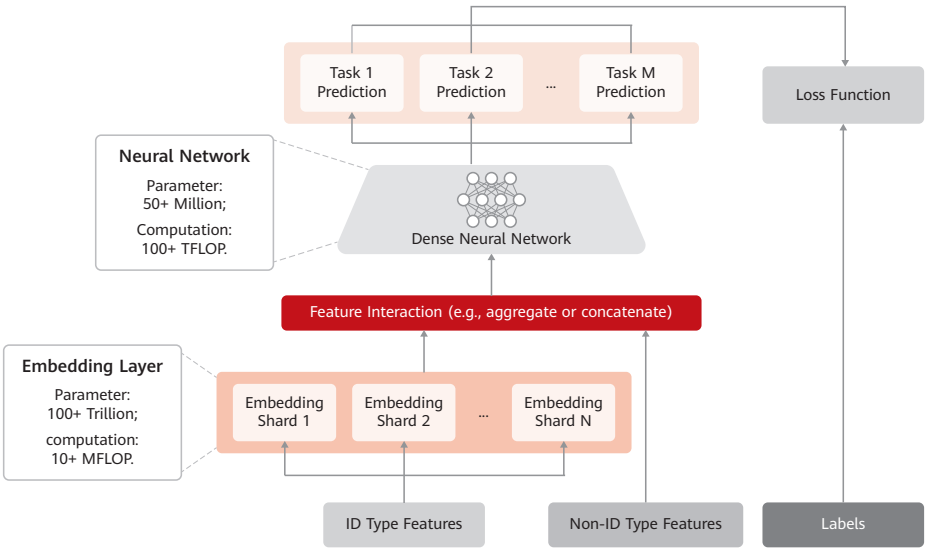


图33 典型的推荐模型结构图

Embedding池化存储：在推荐模型训练过程中，EMS提供全量的Embedding池化存储。另外，在每个训练节点中，本地Embedding模块用于缓存频繁访问的Embedding，并负责与上层推理框架进行交互。当训练节点需要获取Embedding时，首先尝试从本地Embedding缓存中读取。若缓存未命中，则该节点会从EMS中拉取所需数据。在训练过程中，梯度更新任务由数据分片（shard）所属的计算节点执行，并异步将更新后的数据推送回EMS。

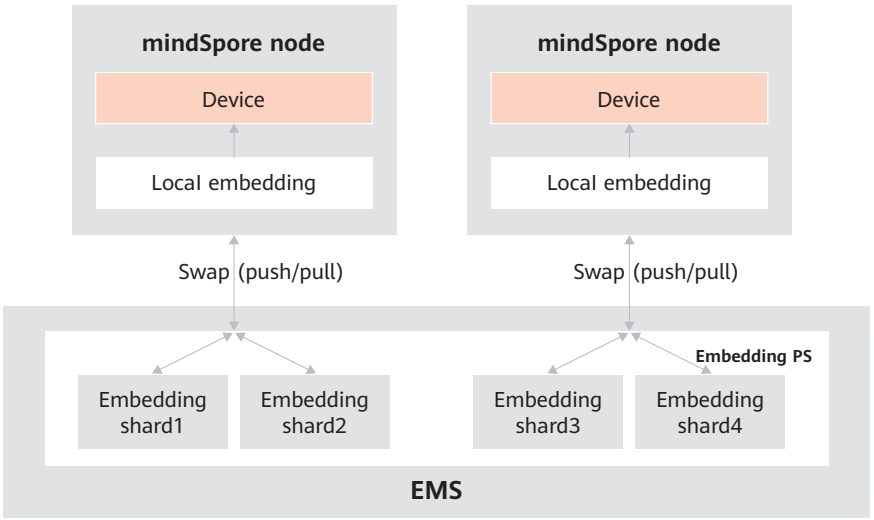


图34 推荐场景使用EMS的架构

- Embedding均衡打散：由于Embedding数据量可能高达十TB甚至百TB，为了提高存取效率，需要将这些数据均匀分散到不同的节点上。EMS采用分片（Shard）机制来分散数据，通过计算键（key）的哈希值，将Embedding数据均匀分配到各个节点。在每个节点内部，再根据哈希值将负载均衡分配给多个分片线程，确保表数据在节点和线程的分片上均衡分布。分片与线程绑定，实现了无锁操作，从而提升了效率和性能。

- 增量检查点：在大规模推荐模型训练中，训练过程中更新的Embedding会被存储到EMS中。在进行训练检查点（CheckPoint）时，也需要保存这些Embedding。EMS提供了增量检查点的功能，只存储更新过的Embedding，而不是每次都存储全部Embedding，这样可以节省存储空间并提高效率。增量检查点的实现基于类似写时复制（COW）的原理，在更新Embedding时记录逻辑时间戳，保存检查点时根据时间戳来判断是否需要保存。

AI场景是EMS的首个应用领域，本书重点介绍了EMS在AI场景中的关键技术。在未来，EMS将持续演进，并扩展至通用计算场景，包括在线事务处理（OLTP）数据库、混合事务/分析处理（HTAP）数据库、向量数据库、Redis缓存系统、大数据分析等应用领域。

» 3.2.5.4 华为云下一代AI原生极简云网络

现有的云网络架构存在以下问题：

- 在数据中心内O/U（Overlay/Underlay）双层复杂组网，虚拟网关转发与运维低效，海量VPC配置生效慢；
- 云广域网设备MPLS多层嵌套协议，带来部署繁琐、运维定位定界困难；
- 跨区域云连接及跨云数据中心按固定峰值预留广域带宽，抬高了租户AI训推、灾备等全局业务成本，且缺乏差异化QoS保障。
- 现有云网络仍遵从纯技术驱动模型，O/U割裂&广域网络/数据中心网络相互独立，缺乏统一协同调度，无法适应应用驱动与租户感知的网络业务流动态按需弹性、灵活路由及端到端精细化QoS保障需求。

针对这些问题，华为云提出了下一代AI原生云网络-CloudGrid极简云网络架构。CloudGrid极简云网络可以更好地支撑AI Native、泛在Serverless及分布式云原生对云网络提出的挑战，从云应用视角出发，对广域网络&数据中心网络进行全面简化和重构，大幅提升云网络端到端性价比、可靠性、运维自动化竞争力。

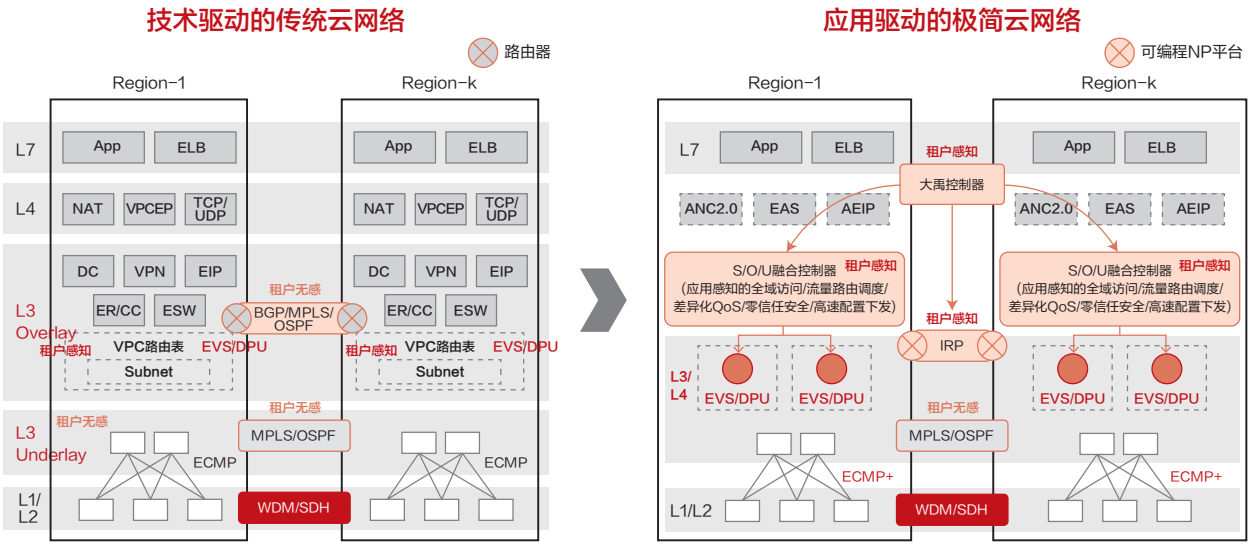


图35 CloudGrid极简网络全景图

CloudGrid极简网络由技术驱动的传统云网络，变为应用驱动的极简云网络，具有8大核心价值：

1 Regionless&跨云弹性网络服务	随着云原生化的场景应用后，服务系统模型越来越复杂，需要的网络互联服务不断增加，尤其是全域跨多云的网络配置越来越复杂，用户网络经验不足，不熟悉云上云上产品，需要熟悉网络多产品使用及运维。ANC（Application Native Cloud，云原生应用网络）提供一种新的网络服务，屏蔽多款网络服务产品配置，全域底层网络随着用户的客户端和服务端的区域位置弹性连通，简化网络连接。
2 ScaleUp总线网络服务化	通过全新数控分离的网络协议栈EAS（Elastic Application Stack）将云网络协议栈暴露的层次提升，直接暴露网络接口。EAS将网络协议栈的带内管控逻辑，与网络数据传输基础逻辑彻底剥离、解耦，即将控制逻辑交给带外管控面，从而实现简化网络数据面的目标，并与云平台的安全、服务等深度融合。
3 无网关瓶颈全域极速互联	当前云服务通过多种网关实现多云互联，跨云连通时需要绕行指定区域位置的集中式网关。在实现了O/U融合的数据面转发后，实现了全网统一编址、消除转发类网关和功能类网关，实现数据面一跳直达。
4 应用全域调度及差异化QoS	全域控制器，拉通数据中心网络控制器和广域网络控制器，实现端到端的租户业务感知和差异化QoS保障。
5 ANC全域应用网格服务	Regionless网络应实现应用像水电一样按需流动，解决算力向低成本区域迁移、低延时多区域就近分发、单区域资源不足阻碍弹性、两地三中心容灾、互联网数据中心生命周期终结等需求。服务访问应提供本地域名固定应用入口，消费者对应用跨区域流动和IPv6改造等流程没有感知。ANC的服务提供全域访问模型，服务一处发布后全域可用，服务IPv6改造消费者无感，企业从容演进。
6 ScaleOut/ScaleUp统一网络模型	CloudGrid提供极简的网络模型，在ANC对象内直接发放AEP（Application Endpoint，应用端点），AEP模型同时支持ScaleOut和ScaleUp域。
7 大规模/高性能按需管控面	在全域网络内高速发放百万规格AEP，且秒级全连通，支撑云原生架构实现高速弹性扩缩。
8 应用级多维度零信任安全	基于CloudGrid网络模型默认隔离，用户无需规划路由和IP地址。通过ANC服务实例的AEP通信，支持服务实例的LadingZone零信任安全，半可信区和核心区通过域策略控制访问策略。

1) CloudGrid极简网络服务模型

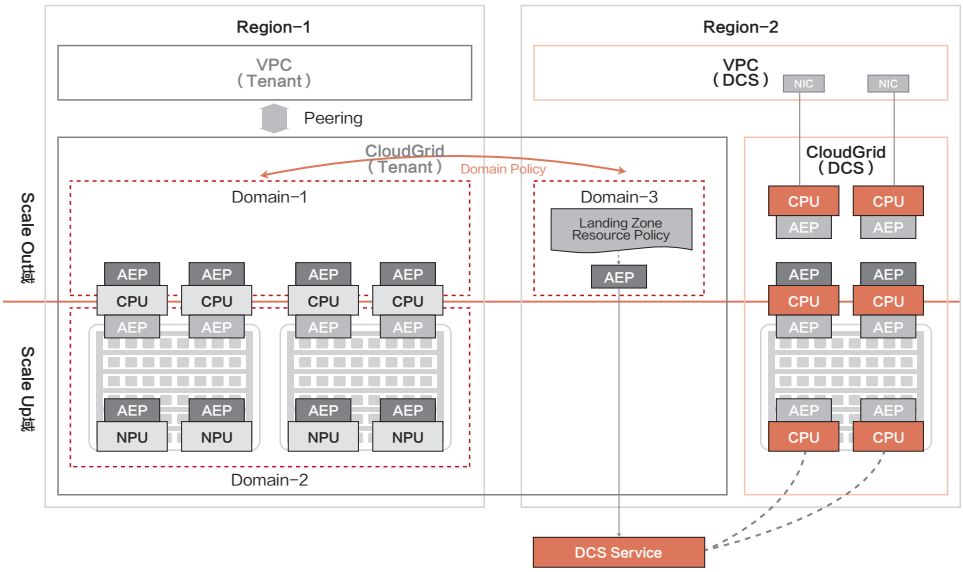


图36 CloudGrid极简网络服务模型

CloudGrid极简网络服务模型包括以下部分：

- 极简CloudGrid模型：全域网络模型（ScaleOut域）：提供ANC对象作为新的全域网络隔离域，无需区域级的VPC/子网模型通过云连接跨区域连通。消除子网及L2转发逻辑：在ANC对象内直接发放AEP，通过AEP即可实现全域直接通信，用户无需感知路由和L2转发逻辑。
- 双栈AEP网络接口：ScaleOut域：支持EAS与IP双栈，实现高性能网络，数据面去网关后一跳直达。ScaleUp域：控制面基于EAS生成转发表项，转发面实现虚拟机/裸金属服务器间高性能通信。
- ANC全局服务：服务多网络协议，全域可见可访问，服务一处发布后全域可用，支持跨账号跨组织发布服务。服务提供域名访问和双栈入口实现IPv4/IPv6互访，服务IPv6改造消费者无感，企业应用可以从容演进。服务后端多区域流转，服务后端可区域部署，支持多类型后端。
- 统一访问控制及安全模型：基于域策略的统一安全策略：基于标签或应用名称的访问控制，而不是基于IP，可以实现安全策略秒级生效和安全规则的超大规格，并支持安全策略全域生效。半可信区和核心区可以通过域策略控制访问策略。零任安全，结合Landing Zone支持多维度参数。EAS管控面根据域策略、服务资源策略，实现双端固定数据边界安全。通过统一上下文，实现多维度零信任访问策略保护。

2) CloudGrid广域网络智能路由协议IRP

IRP（Intelligent Routing Protocol）：智能路由协议。CloudGrid广域网络部署采用SDN网络架构，IRP转发节点采用可编程设备实现，部署在华为云边缘云和中心云节点；IRP中心控制器通过综合多因子选路，为业务选择最优路由下发给IRP转发节点，从而实现业务的全球极简端到端QoS的传输保障。

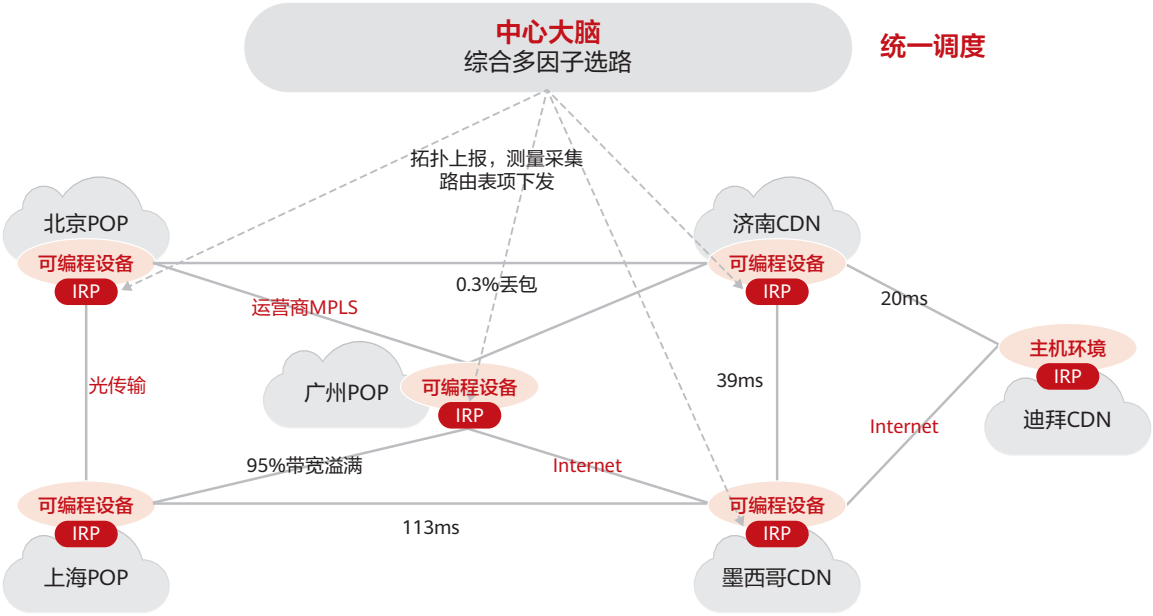


图37 CloudGrid广域网络部署组网图

IRP极简广域网络优势1: 协议极简，去MPLS VPN多层复杂协议，极简部署与运维。部署了IRP协议后，可以实现千级节点协议极简化，单节点部署从1小时减少至10分钟。

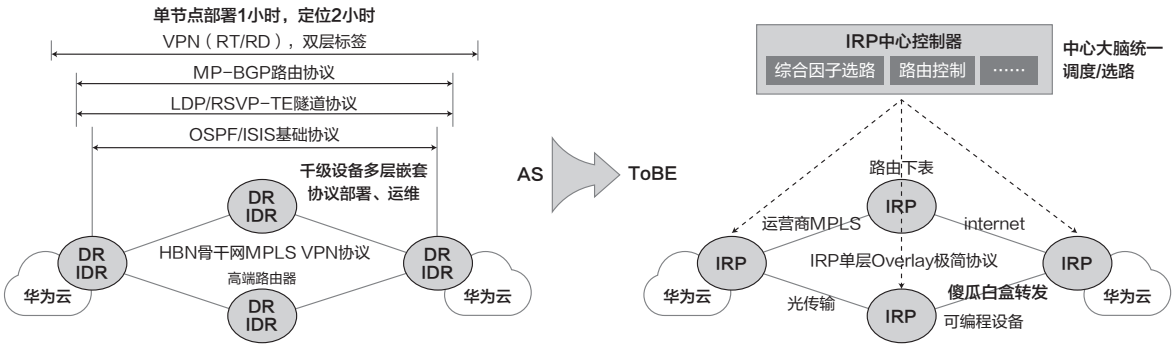


图38 IRP协议极简

IRP极简广域网络优势2: 东数西算，低成本组网，共享弹性带宽，降低Regionless广域传输带宽成本。

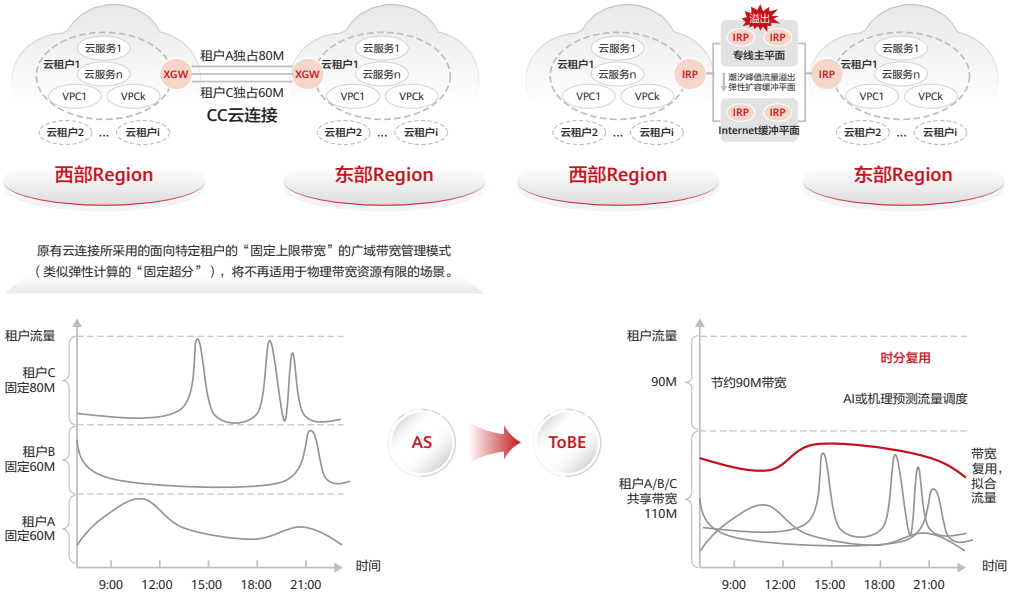


图39 IRP降低传输带宽成本

- “东数西算”网络当前采用云连接方式，云连接为租户分配固定带宽，存在潮汐闲时浪费，造成带宽浪费。
- 采用IRP广域网络承载后，多租共享弹性带宽，消峰错谷；主用专线平面，使用AI流量画像，在流量溢出时平滑切换至冗余Internet网络平面。

3) CloudGrid极简网络分层控制器

当前业界数据中心网络控制器和广域网络控制器是割裂的，没有一个统一控制器来进行全局访问控制和流量调度，以及租户的端到端业务体验保障等。CloudGrid极简网络首次提出分层控制器的方案，即有一个全局控制器，拉通数据中心控制器和广域网络控制器，形成全局一张网，其架构如图40所示：

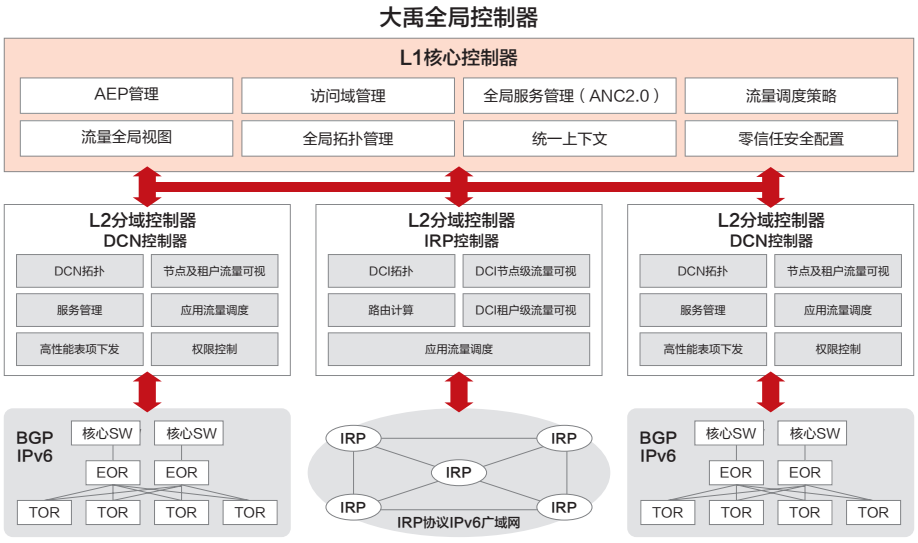


图40 大禹控制器架构图

- L1层的大禹控制实现了全局统一管理，包括全局拓扑管理、流量调度、访问控制等；
- L2层的数据中心网络控制器/IRP控制器实现了分域高性能控制，包括区域的拓扑管理、表项下发、流量调度等；
- 两层控制器协同实现了全局节点&租户级流量可视以及端到端差异化QOS服务保障。

» 3.2.5.5 柔性计算多元算力范式

随着日新月异、不断取得一个又一个前沿突破的大模型、智能体技术在千行万业的深入落地并转换为生产力提升，AI算力需求量呈现出爆炸式增长的趋势，使其成为数字经济时代像水和电一样不可或缺的最核心、最基础的生产资料。同水和电一样，支持一键式弹性按需获取并按租赁时长计费的云上AI算力，提供了相比线下AI算力更便捷高效、更优性价比的算力供给模式。鉴于此，华为云提出柔性智算，通过对昇腾不型号的NPU实现用户态虚拟化（FlexNPU）、AI驱动的智能伸缩与混部、AI模型算力建模画像以及AI算力全域多优先级抢占式调度4大关键创新，旨在大幅提升昇腾云AI算力集群的利用率水平、弹性伸缩能力及可靠性可用性SLA，并实现对云上千模百态的大小模型，以及业界主流AI训练与推理框架的广泛兼容，打造面向昇腾云的下一代Serverless AI算力基础设施新范式。具体关键技术细节将在本章节展开。

1) 柔性智算的整体架构

华为云柔性智算通过将柔性计算的核心理念与设计原则从通用计算延伸到智能计算领域，使能模型训推算力动态需求感知的细粒度AI算力分配调度与弹性伸缩，并通过训推极致混部、推理PD动态混部、与A3/A5超节点UB网络深度协同等差异化创新，重新定义“云上AI算力”在极致性价比、Serverless化弹性伸缩和高可靠高可用性方面的业界新基准。

柔性智算的整体架构设计如下图所示：在当前业界主流AI开发框架与昇腾NPU CANN驱动软件层之间，引入了“NPU用户态虚拟化（FlexNPU）、AI驱动智能伸缩与混部、AI模型算力建模画像、AI算力全域多优先级抢占式调度”4大创新技术引擎，实现了云上运行的百模千态的AI训推任务所感知的“虚拟NPU算力”与部署在全球AI云数据中心内的“物理NPU算力”之间的解耦与灵活映射，在满足AI训推任务性能SLA的前提下，最大限度提升昇腾云NPU算力集群的总体利用率，降低其无效空转的比例，并有效屏蔽昇腾NPU硬件故障对上层训推任务及AI框架软件层带来的可用性与业务连续性影响，从而打造面向下一代的AI训推框架透明、极致弹性&利用率、极致高可用的Serverless AI算力底座。

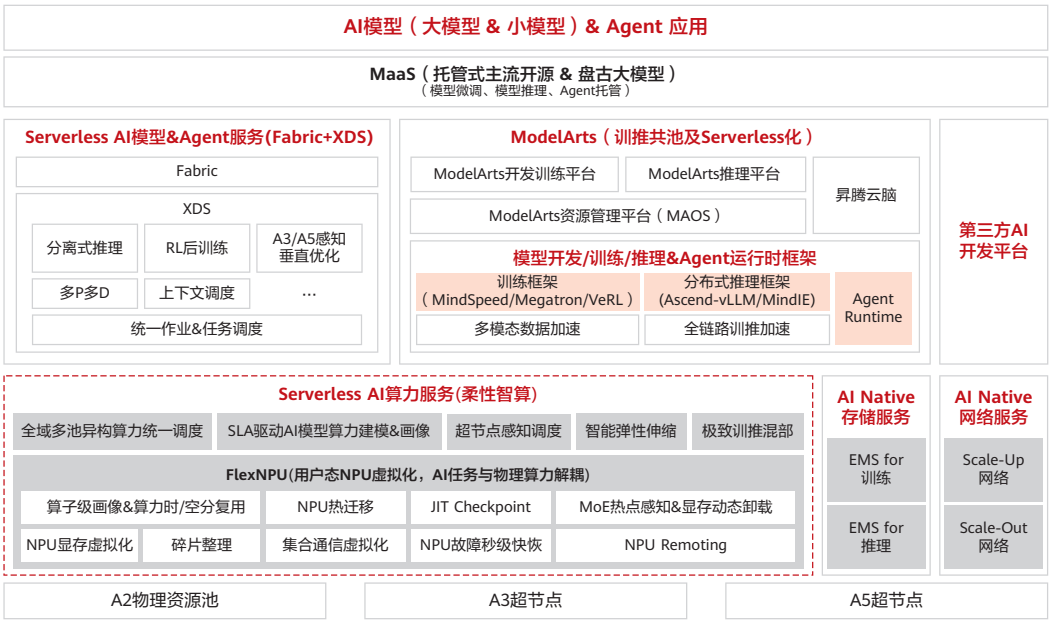


图41 柔性智算整体架构

- NPU用户态虚拟化（FlexNPU）：该技术引擎透明拦截训练及推理任务经主流AI框架（Pytorch，MindSpore，vLLM等）对昇腾算子的CANN API调用，并参照NPU算力池内每张NPU卡在AI Core、显存、带宽等各个维度的实时资源利用率与忙闲状态，各AI模型对应的虚拟NPU规格，以及当前下发算子的执行时长与算力规格预估，按照一定的时分及空分调度策略，将上述昇腾算子的CANN API调用合理有序地分配到多模型共享的物理NPU卡上，从而达成在满足AI训推任务性能SLA的前提下，多个AI训推任务对NPU硬件算力的最大化空分与时分复用；与此同时，该技术引擎还支持AI训推任务及框架无感，CPU与NPU协同，以及多NPU卡协同的事务一致性运行时快照，从而在无需云上租户及其AI训推任务干预和介入的情况下，将运行中的AI训推任务，以最短业务中断时长为代价，从受NPU硬件或驱动软件升级影响的物理节点，迁移到备用或空闲的物理节点上，或者从碎片化整理的特定源节点迁移到特定目标节点，将上述重要的例行运维活动对云上训练和推理任务的业务可用性与连续性影响降到最低；更进一步，针对超节点上的大模型并行推理和并行训练场景，该技术引擎还支持主动拦截和屏蔽昇腾NPU硬件上报的单点故障事件，并通过超节点内跨节点的NPU Remoting机制，实现N+1备用节点上的备用卡对故障卡的秒级快速透明接管，防止故障爆炸半径的扩散的同时，将单点NPU故障对训推任务的RTO影响从几十分钟降低到秒级水平。
- AI驱动的智能伸缩与混部：考虑到所有推理服务一般均具备显著的“潮汐效应”，该技术引擎基于大颗粒在线推理服务的历史算力资源用量及业务请求量统计数据，以及基于Transformer架构的多维度输入AI时序预测基础模型，面向不同MaaS推理服务进行可持续迭代的智能伸缩预测模型SFT后训练，并基于该模型指导AI推理服务进行容器实例及其节点的预测式扩缩容，用以替代传统基于特定监控指标阈值的规则式弹性伸缩机制，从而在保障在线推理服务性能SLA的基础上，以合适的时机及步长进行在线推理AI算力的申请和释放，并支持将相关算力释放给其他训练与离线推理任务，从而打破多推理任务及训推任务之间的算力孤岛，实现极致的AI算力共享。
- AI模型算力建模&画像：为实现公有云数据中心内数以十万计乃至未来还将不断增长的NPU卡存量算力资源供给，与来自云上租户多种模态、不同尺寸AI模型的多样化算力需求之间的最优匹配，宏观层面上多用户动态叠加的AI总体算力量化建模和画像对公有云AI算力服务运营效率及投入产出比的优化提升具有重要意义，鉴于多租MaaS服务已成为昇腾云超节点算力的主力场景，如何基于MaaS推理业务层指标（如RPM/TPM等）建模仿真能力，指导MaaS服务算力池进行合理步长的容量规划，就成为当前昇腾云宏观算力建模的焦点问题。而除此之外，与NPU用户态虚拟化（FlexNPU）技术相结合的细粒度微观层面的AI模型算力建模能力也同样不可或缺：特别是针对多个小模型共NPU卡的时分和空分混部，以及大模型在特定并行策略下NPU卡碎片化资源与其他AI模型的共卡混部，都需要AI算力建模工具具备依据AI模型“计算图及算子”解析进行细粒度多维度算力量化估算评估的能力，从而为FlexNPU进行物理NPU到虚拟NPU的合理切分提供关键输入和依据。当然，考虑到软件的迭代优化及NPU硬件代次演进的影响，也有必要结合AI模型及算子级的实际算力用量情况的精细化观测与画像，对AI模型算力的理论建模结果进行必要的修正。
- AI算力全域多优先级抢占式调度：为解决当前云服务商的AI算力普遍采用Region内调度所带来的跨Region供需不均，以及租户独占AI算力所导致跨租户动态忙闲不均等关键痛点，柔性智算构建了跨云内所有Region的统一AI算力视图，以及AI训推任务SLA驱动的自动化全域AI算力调度能力，使得昇腾云AI算力可以在云内实现真正的跨租户、跨区域最大化共享，并且全域调度的范围不限于AI算力的初始调度，也包括因高优先级AI任务抢占式调度所触发的中低优先级AI任务基于FlexNPU透明快照的跨Region二次调度。

柔性智算创新对昇腾云的核心价值：

- 昇腾云算力有效利用率大幅提升，昇腾云应对友商AI算力价格竞争拥有更充裕的利润缓冲空间：通过上述柔性智算4大创新技术引擎的构建与推广，满足云上多用户百模千态的AI模型训练与推理任务在特定性能SLA约束条件下的NPU硬件算力总支出将大幅降低，从而为昇腾云从容应对来自其他云服务提供商，特别是GPU算力服务的白热化价格竞争，提供足够的利润缓冲空间。
- 柔性智算带来的NPU算力性价比、弹性及高可用提升等核心价值，可同时覆盖基于MaaS的Token服务场景，以及面向NA大客户训推任务的昇腾算力托管及租用场景：基于FlexNPU的小模型时分/空分共卡、大模型轻量化、极致高可用（含AI任务无感的冷/热迁移，训练推理任务的秒级快恢等），以及极致训推弹性混部等能力，广泛兼容PyTorch、MindSpore、Tensorflow等主流深度学习框架，以及vLLM、MindIE等主流推理平台的昇腾适配版本，因此可普适应用于昇腾云基于MaaS的Token服务场景，以及面向内外部NA大客户的昇腾算力托管与租用场景，含Lite Server、Lite Cluster以及Standard Cluster以及HCS混合云等场景。
- 支撑MaaS服务的每Token性价比更上一层楼：通过柔性智算面向多租MaaS服务在宏观层面的AI模型算力建模所带来的超节点算力容量整体规划效率的提升；基于FlexNPU透明快照机制+AI弹性伸缩预测模型的多租户MaaS服务之间的动态混部，MaaS推理服务与其他推理业务之间的细粒度混部；基于FlexNPU时分/空分复用虚拟化+微观层面AI算力建模的MaaS推理PD动态混部，推理性能SLA驱动的显存动态卸载，以及MaaS推理任务与其他多租小模型之间的NPU卡复用，也将支撑相同NPU硬件算力投入及推理性能SLA约束前提下，达成更大的Token吞吐率，从而推动昇腾云MaaS服务的每Token性价比再上一层楼。

2) 柔性智算关键子系统及其核心技术

i. 面向多模型共NPU卡的时分与空分虚拟化

基于FlexNPU用户态虚拟化技术的AI算力灵活切分与复用是柔性智算数据面最核心的子系统之一。它解决了AI训练与推理框架以NPU物理卡作为训练与推理进程的AI算力最小单元的“粗放式”资源分配模式所带来的算力浪费问题，该技术通过AI Core时分/空分复用与显存虚拟化两种机制，分别实现对计算资源和内存资源的细粒度管理，数据面技术架构及关键技术如下：

- AI Core时分复用：基于算子级细粒度预画像，实现了多模型共卡混部场景下算子级AI Core的时分复用。系统通过精确掌握每个算子的执行特征，智能调度不同模型的算子任务，最大化利用计算资源。与NPU卡硬切分的空分复用方案相比，此项技术主要针对有明显潮汐特征的AI任务进行混部，时分复用技术能在保障AI任务性能SLA的前提下，能实现AI算力的最大化利用。具体实现中，调度器将时间划分为微小的时间片（通常为毫秒级），根据不同任务的优先级和特性动态分配时间片，确保高优先级任务获得及时响应，同时保证系统整体吞吐量。
- AI Core空分复用：FlexNPU借助CANN层进程级别的Device资源限制接口，基于AI任务声明的AI算力需求，对多任务共卡混部时使用的AI算力进行限制，确保AI Core的QoS隔离；与NPU卡硬切分不同的是，此切分能力为按需的软件层切分，避免了预先固定切分的资源浪费。此项技术主要针对多个特别小的模型共卡混部时，基于AI Core的空分复用提升系统吞吐并保障性能。

- NPU显存空分复用: FlexNPU通过对昇腾显存操作相关算子的透明拦截处理，实现了多AI任务共卡的显存资源空分复用所需的隔离性、透明性，以及虚拟物理显存地址建的高效转换。该技术通过引入显存虚拟地址空间，使每个AI任务拥有独立的显存视图，互不干扰。同时，虚拟化层负责物理显存资源的分配与回收，以及虚拟地址到物理地址的转换，确保内存访问的安全性与高效性。
- NPU显存时分复用: 针对推理请求频率较低的长尾模型，FlexNPU不仅支持多个LLM模型显存之间的低延迟“时分复用”，也支持基于各LLM模型的活动性及动态处理容量预测、H->D显存Prefetching换入以及基于模型PP切分的推理与加载流水线并行机制的多模型共卡的显存时分复用机制，从而支持在多模型推理请求存在显著的热点与长尾两极化特征情况下，可在满足推理性能SLA前提下，大大提升平均每NPU卡的模型复用比（从平均每NPU卡复用2-3 LLM模型，提升至每NPU卡复用7-8 LLM模型）。

ii. 面向大模型轻量化部署的MoE热点专家感知的显存动态卸载

业界主流的开源MoE大模型（如DeepSeekV3，Kimi2，Qwen3等）在公有云行业专属隔离区，以及CloudPond专属边缘部署场景等下，每Token的实际激活专家数仅占MOE模型总专家数的3.5%，但仍需所有MoE专家100%常驻NPU显存，从而导致MOE大模型的本地化部署成本始终高于用户预期的典型痛点（通常基于昇腾A2节点，W8A8精度的DeepSeek V3/R1典型配置需要2台313T或4台280T），FlexNPU着力构建了“热点专家感知的显存动态卸载”差异化竞争力: 使得业界主流的千亿/万亿级参数开源MoE大模型的W8A8典型硬件配置从2台313T昇腾A2节点缩减到1台，而其W4A8典型硬件配置也从最小一台313T昇腾A2节点推广至一台280T昇腾A2节点。

支持仅MoE热点专家常驻 HBM，长尾专家则常驻主机内存，或按需动态从显存换入内存。而当系统重新需要长尾专家时，则支持将对应的长尾专家权重通过 A2 节点的DMA/PCIe 通道，或A3超节点的HCCS链路精确从内存拷回显存；配合 LRU + 黑名单维护热点集合，动态换入延迟压缩到百微秒量级，对上层模型完全透明。

卸载管线包含“识别 → 驱逐 → 回映”三步。调度器在GEMM启动前结合TopK路由和访问统计做在线打分，维护热点专家集合；若命中专家不在HBM，则触发驱逐策略写回冷门专家，并从主机内存并发回拷目标专家；若专家仍在HBM，仅刷新热度即可。这样既保证了执行路径畅通，又避免了长尾专家长期占用显存。

为确保推理性能SLA的满足，FlexNPU进一步引入了“二次路由”增强: 在门控输出的 TopK 基础上，参考 HBM 当前分布做一次轻量 reroute，使路由尽量命中已在显存中的专家，减少搬运次数。通过用户定义的SLA，动态调整路由换入策略。

进一步的大模型本地化部署的性能SLA深度优化措施包括:“提前预装载领域热专家 + 预测预取 + 多路径传输”: 在NPU计算的时候，结合预测实现主动预取，提前触发下一层的H2D的专家搬运，掩盖搬运时间，同时利用分桶路由存储的方式，缓解某device专家不均衡的问题，打造面向超大规模MoE集群的显存治理能力。

iii. AI模型算力建模&画像

华为云面向众多租户提供、特定版本的开源及自研AI模型，无论是需要多NPU卡、多节点支撑的千亿/万亿参数规模的大模型，还是单NPU卡即可容纳的百亿级以下参数规模的小模型，由于其模型架构、计算图，乃至构成计算图基础节点的每个算子都是已知和确定的，因此完全可通过AI算力建模工具或服务对其算力需求提前进行精细化地白盒式建模测算，并在模型上线运行后进一步通过黑盒式的在线资源画像，对其在特定代次昇腾NPU硬件上实际算力消耗进行洞察，以便对理论测算值进行最终的校准与修正，而这些白盒建模及黑盒画像的AI模型资源量化需求，即可在宏观层面商为公有云服务提供商的AI算力总体需求规划提供核心依据，也可在微观层面为FlexNPU控制面的全域调度及数据面的时分及空分复用，提供关键输入。

首先看微观层面的AI模型算力量化建模：其核心思路是通过“计算图+算子”的白盒式NPU算力需求解析建模方法，以AI模型训练与推理的性能SLA为前提约束，为AI模型自动推荐最优的细粒度资源配置（精细化的AI Core及HBM显存容量），并为大模型推荐优选的多维度并行策略。

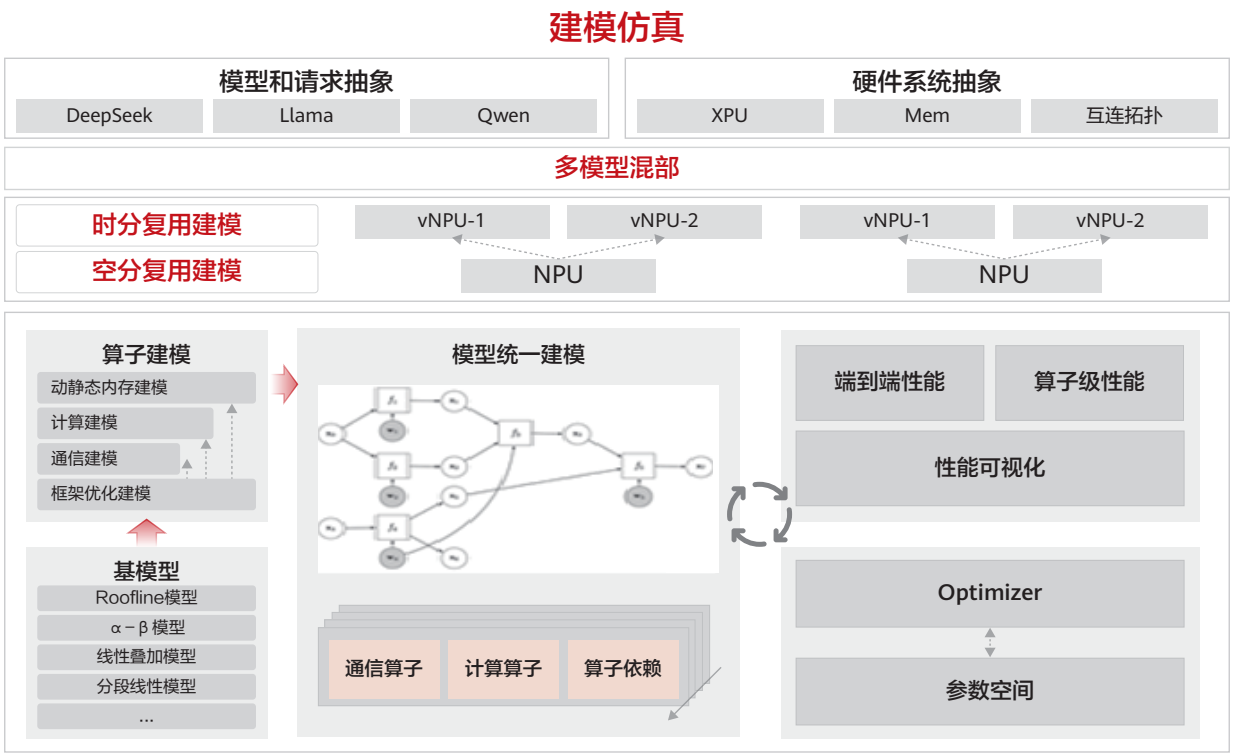


图42 AI模型的微观算力量化建模

MaaS（Model-as-a-Service）平台在运营成本上面临的核心挑战是如何在保证推理SLA的前提下，实现多租MaaS服务用户所共享的MaaS超节点集群的资源高效利用。传统方法会根据租户的诉求为每个租户单独设置预留资源量，并在推理服务的整体资源预留上设置一个超分比。这一方法存在两个问题，一是用户很难准确评估自己的资源诉求，二是超分比的设置缺乏理论依据，通常是依据用户经验和多次试错逐步收敛至合适值。

对此FlexNPU引入了概率叠加容量规划算法，通过分析租户流量的统计特征，优化整体资源分配，避免简单的峰值叠加导致的资源浪费。此算法将集成到璇玑平台，并指导MaaS服务进行服务级的资源调整。

简言之，通过将柔性计算的概率叠加算法引入到AI算力峰值容量预测场景，解决了各MaaS托管服务并发推理请求数（及其对应的AI算力资源需求量）在特定时间点的取值难以直接进行确定性估算的关键挑战，通过每个托管MaaS服务并发请求的历史时序数据转换为不同并发请求数区段（及其对应的AI算力资源需求量区段）的确定性概率分布，巧妙地将AI算力容量测算问题转换为独立随机变量之加和的概率学基础问题，达到了对昂贵的AI超节点算力整体资源峰值更为精准的评估预测。

iv. AI模型训练与推理任务故障秒级快速恢复

FlexNPU虚拟化层通过透明拦截AI训练和推理框架的NPU算子调用，拥有全量的NPU算子的API调用记录（Redo log），基于该特性构筑了与昇腾云AI算力解耦的训练与推理任务的跨节点冷/热迁移的能力，使得公有云主动运维场景下的NPU硬件更换/升级、NPU驱动/CANN软件重大版本升级，以及NPU算力资源碎片化整理等对AI训练与推理任务的业务可用性、连续性有重大影响的运维管理活动，也能像通算场景下的冷/热迁移那样，对云上的NA用户做到基本透明无感。

除此之外，FlexNPU虚拟化技术还可支撑昇腾云更有效地解决偶发的NPU单点故障的爆炸半径因TP/PP/DP/EP等模型并行策略而被扩大到整个超节点/整个集群，并进一步导致训练与推理任务的故障恢复RTO过长（数十分钟），以及推理KV Cache上下文丢失等核心痛点问题。在单点故障场景下，由FlexNPU用户态虚拟化层对NPU硬件返回的部分报错信息进行了拦截处理（部分由用户造成的错误透传至用户侧），在训推任务用户无感情况下实现了一系列故障恢复动作，并且由于系统级快照中已包含NPU稳态上下文而省去了AI框架初始化过程，可基于NPU算子调用的Redo log重放恢复NPU的动态增量上下文状态，最终有力支撑了大模型训练及推理任务的RTO时长从数十分钟到数秒的优化。

v. AI驱动的智能弹性伸缩及极致训推混部

在线MaaS推理业务请求及其AI算力需求，往往呈现显著的潮汐特征，通过将AI训练及推理业务混部到同一算力集群，充分利用启停时机更灵活、离线作业型的AI训练业务来填充在线作业型的AI推理业务在非忙时段空闲下来的AI算力资源空间是比较常见的AI算力资源效率优化手段。然而，当前AI训推业务的混部普遍存在3大关键挑战：一是缺乏对推理业务总量在未来一定时段内精确量化的变化趋势缺乏准确的感知，因此往往只能选择采用基于经验规则的预定义阈值触发的方式进行推理集群的扩缩容管理：若预定义阈值过于保守，则将导致多数时段存在显著的AI算力浪费；而若预定义阈值过于激进，则将导致由于AI算力就绪不及时，推理请求被迫排队等待，从而影响业务SLA的满足度；二是针对大模型训推混部场景，训练和推理业务及其关联数据从同一AI算力节点换出及换入的“中转”耗时较长，特别是从训练业务切换到在线推理业务（在线推理业务对时延更为敏感），AI算力节点上的存量训练业务需要先暂停并做完快照，才能启动推理框架及AI模型参数及其KV Cache数据从CPU到NPU的串行加载与初始化流程，上述长达数十分钟的“中转”时段内的AI算力资源无法及时为高优先级的在线推理服务提供支撑；三是针对小模型训推混部场景，混部在同一NPU卡上的训练与推理任务，由于缺乏细粒度的算力分配与隔离机制，且硬件上下文切换代价高，导致无法保障对训练和推理各自性能SLA的遵从和满足。大模型训推混部的场景介绍架构如下图所示：

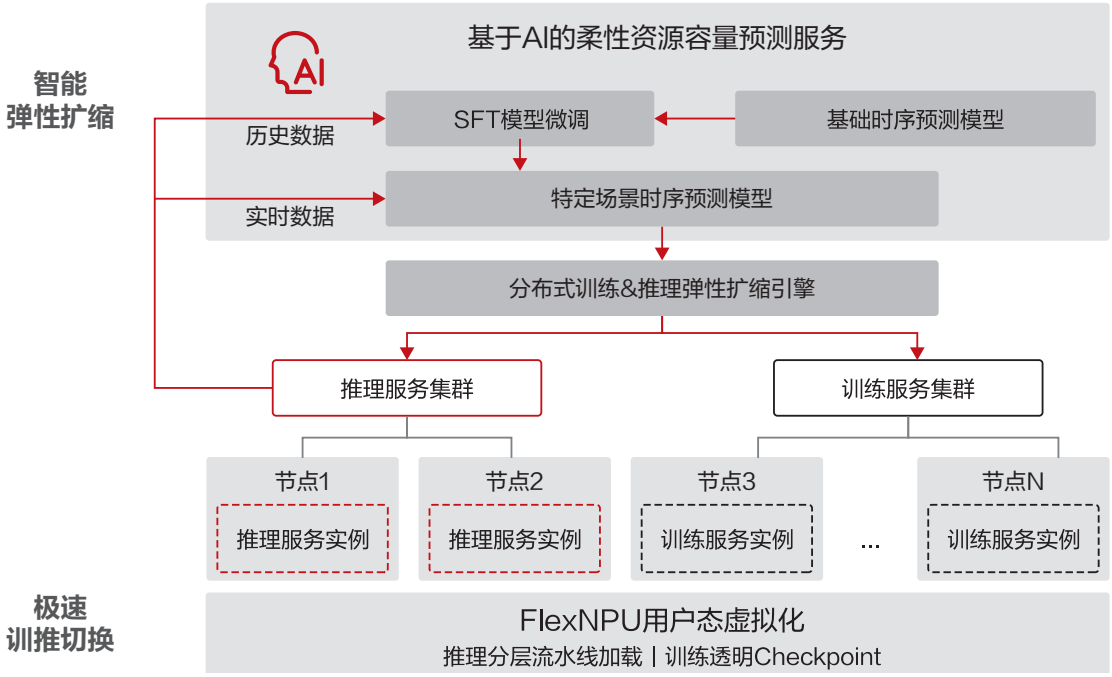


图43 基于FlexNPU的大模型训推极致混部架构

大模型推理服务和训练服务以节点为粒度切分底层的共享算力资源，推理服务动态伸缩Prefill和Decode实例的数量，训练服务则根据推理服务的容量动态调整DP分片数，训练和推理任务同时运行，但不同时间段资源占比不同。

在控制面，通过引入AI驱动的智能伸缩机制，基于推理服务的历史流量数据，基于Transformer架构训练时序预测MoE基础模型，进行匹配当前推理服务的AI智能伸缩模型的SFT训练，基于该时序预测微调模型，以及近期1小时时间窗内的AI推理业务与算力用量统计数据作为模型输入进行预测推理，既能完美解决固定预设弹性伸缩阈值设置过于保守所带来的AI算力闲置浪费，也能保障在实际推理流量高峰到来之前，及时进行合适步长的AI算力水平扩容与发放，避免了业务突发高峰期因AI算力制备不及时导致的忙时业务等待。

依据从昇腾云数据湖导出多租户MaaS推理平台各服务的推理流量数据，并应用上述Transformer的AI驱动弹性伸缩时序预测模型，对训推混部的空间做了评估。从上面推理服务平台总RPM实际和预测曲线可以看出，推理流量的波峰波谷之间有着显著的资源空间。根据各服务推理流量的峰值、均值，以及服务的实例数和卡数，并考虑FlexNPU的AI容量预测和用户态虚拟化带来的训推切换及时性提升，预计当前昇腾云的MaaS推理平台可通过AI驱动的推理算力水平伸缩挖掘用于支撑训练任务混部的NPU算力空间，约占当前MaaS推理服务总算力的34.87%。

vi. 全域AI算力多优先级抢占式调度

提升云上NPU算力利用率的关键措施就是打破算力孤岛，促使算力在多个部门和业务之间极致共享。FlexNPU构建业务层面的虚拟NPU配额管理机制，使业务团队与物理NPU设备解耦，支持NPU设备在多个业务之间快速共享，通过引入多优先级AI任务调度机制，支持高优先级任务优雅的抢占低优先级任务所使用的资源，同时借助跨区域调度能力，将低优先级任务迁移至其他满足算力条件的Region运行，实现了算力的高效共享。具体架构图和关键技术如下所示：

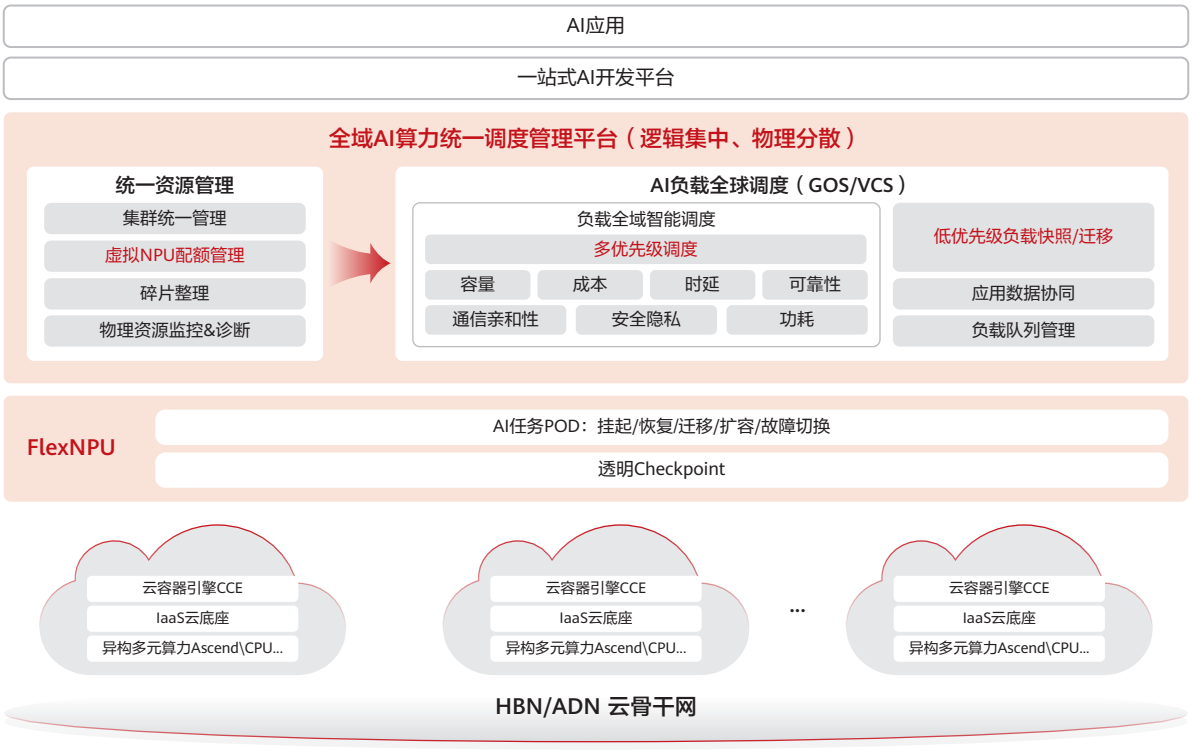


图44　全域AI算力多优先级抢占式调度

- NPU算力配额虚拟化：核心在于打破“一个团队独占一批物理GPU”的传统资源隔离模式。通过将全球数据中心的NPU资源整合为一个统一的虚拟化资源池。各团队获得的不再是物理算力单元，而是虚拟化的算力配额。这种模式从根本上解决了资源孤岛问题，使得一个团队未使用的配额可以立即被其他团队的任务动态共享，从而为实现全局范围的资源超售和超高平均利用率奠定了基石。
- 多优先级抢占式调度：为确保关键任务能即时获取算力，平台引入了三级优先级（低、标准、高）的抢占式调度机制。当高优先级任务需要资源时，它可以抢占正在运行的低优先级任务所占用的NPU。然而，其关键创新在于“抢占”并非“终止”，而是一种受控的驱逐。调度器保证被抢占的低优先级任务不会被直接杀死，而是进入一种等待重新调度的状态，从而在保证系统整体吞吐量的同时，也兼顾了不同优先级任务的公平性与完成进度。
- 跨区域冷迁移：这是实现高效抢占和资源复用的关键保障技术。当某个区域集群内的低优先级任务被高优先级任务抢占后，全局调度器不会简单地将其挂起等待本地资源空闲，而是会主动在全局资源池中寻找其他区域的空闲NPU，并将被驱逐的任务重新调度（冷迁移）到目标区域启动。这个过程对于任务本身是透明的，虽然会有关机再启动的开销，但确保了任务最终能够完成，从而将原本可能被浪费的“驱逐”事件，转化为一次跨地域的资源再平衡操作，极大提升了整个资源池的韧性和效率。

基于上述全域AI算力的多优先级抢占式调度机制，可进一步实现不同租户跨Region的训练与推理任务之间，以及在线与离线推理任务之间的“抢占式调度”，从而确保高优先级任务可通过抢占中低优先级AI任务优先获得AI算力资源，而被抢占任务，也可通过充分重用云内所有Region当前处于闲置状态的碎片化AI算力资源完成AI算力的二次重调度，以便继续从快照断点恢复继续运行其任务。

全域训推混部在更大的范围进行算力资源共享，当一个区域内的训练任务被推理服务抢占，该训练任务无需等待该区域腾出空闲资源再继续执行，而是先基于NPU虚拟化先进行任务快照，再通过全域AI算力统一调度系统寻找其他区域的合适节点，将该训练任务冷迁移过去，并恢复运行。跨区域的训推混部，可以更好的保障训练任务的执行效率，也能够在区域间进行资源用量平衡。此场景的架构图如下：

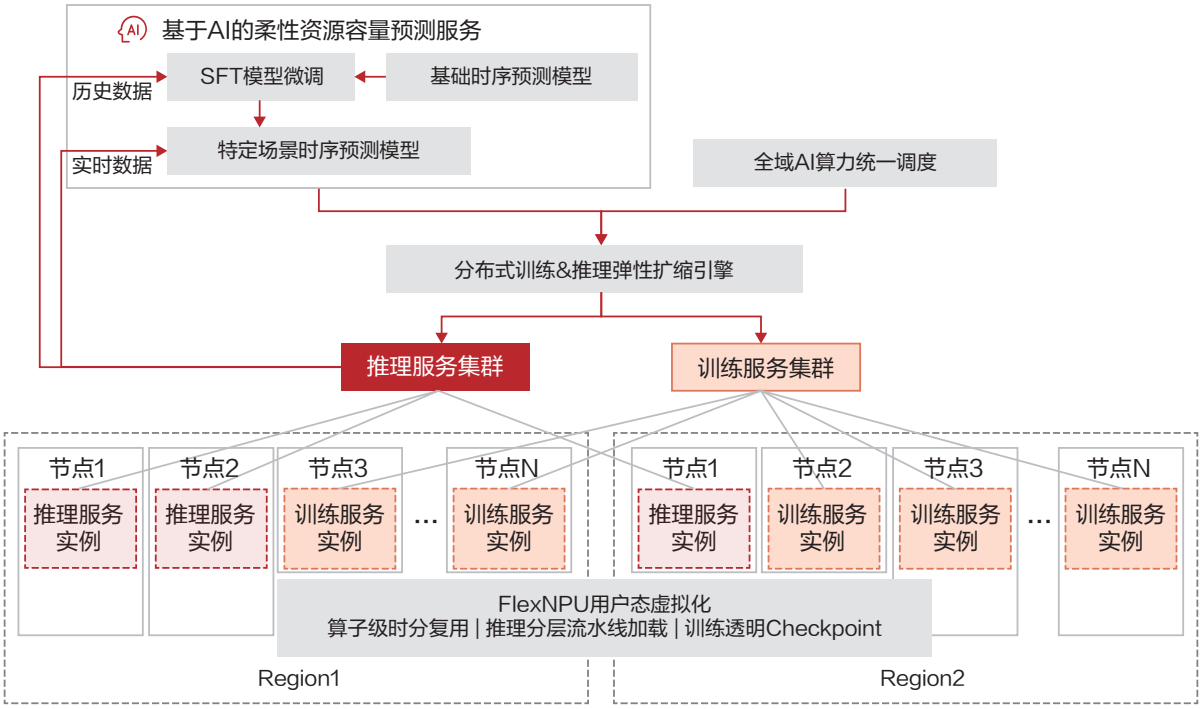


图45 全域拉通的训推任务抢占式混部与调度

» 3.2.5.6 昇腾云算力平台及生态迁移工具

随着人工智能技术的飞速发展，我国在诸如科学研究、工业制造、智慧城市、智慧医疗等众多领域对AI算力有着巨大且不断增长的需求，算力已成为推动其进步的关键因素。为支撑国内AI产业生态，AI软硬件技术近年取得了长足的发展。同时，为避免跨域数据传输带来的安全与隐私问题，促进国内人工智能技术的健康发展，在当前阶段构建算力资源管理平台和生态迁移工具链具有极其重要的价值。

1) 资源管理概述

作为一个完整的AI平台，需要具备AI训练任务管理，推理任务管理，AI开发环境等AI相关能力，而为了让训练子系统和推理子系统可更好的聚焦AI训推性能优化，将两者的作业任务进行统一的管理是一个较好的实践。因此需要有一个独立的任务管理及资源调度子系统，来简化上层训推业务的开发。

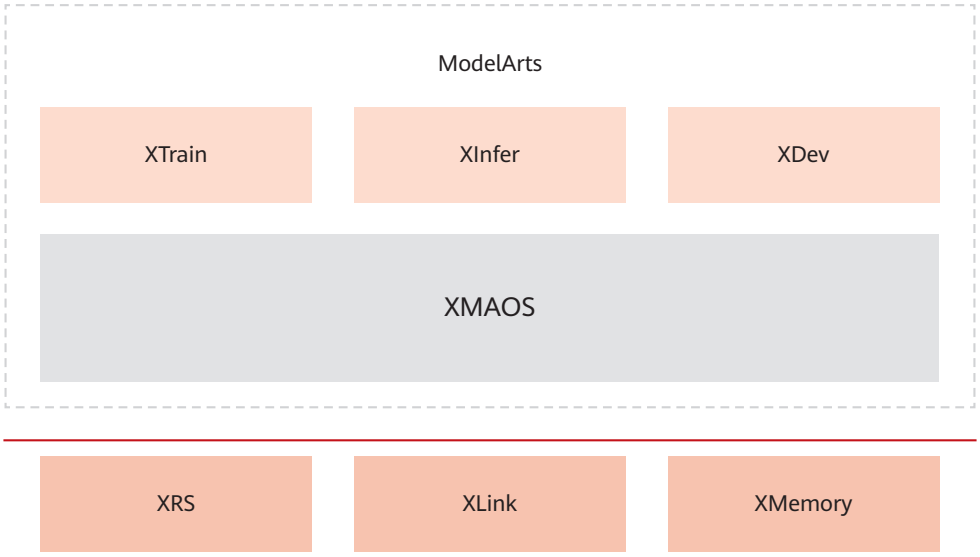


图46 昇腾云算力平台资源管理上下文架构示意图

MAOS-MA作为整个云系统AI算力的管理系统，主要承担上述这个职责：（1）首先是资源的管理。负责管理所有NPU资源以及NPU的购买发放。（2）其次是资源的调度。为AI任务调度分配合适的NPU卡，确保分布式AI训练或者AI推理，调度到可用的NPU节点。MAOS负责资源管理+任务调度，并支撑训练+推理子系统的业务。其中统一任务包括各类训练任务、推理任务及开发任务。通过与训练+推理子系统的分层构建，使它们能聚焦训练+推理本身的优化。

此外，由于故障在AI集群中是不可避免的，因此在任务的资源调度分配后，MAOS还需要确保训练任务的稳定运行。即在故障发生后，分布式训练任务可以快速的恢复并利用checkpoint重新进入训练，因此任务故障重调度也是MAOS职责的一部分。

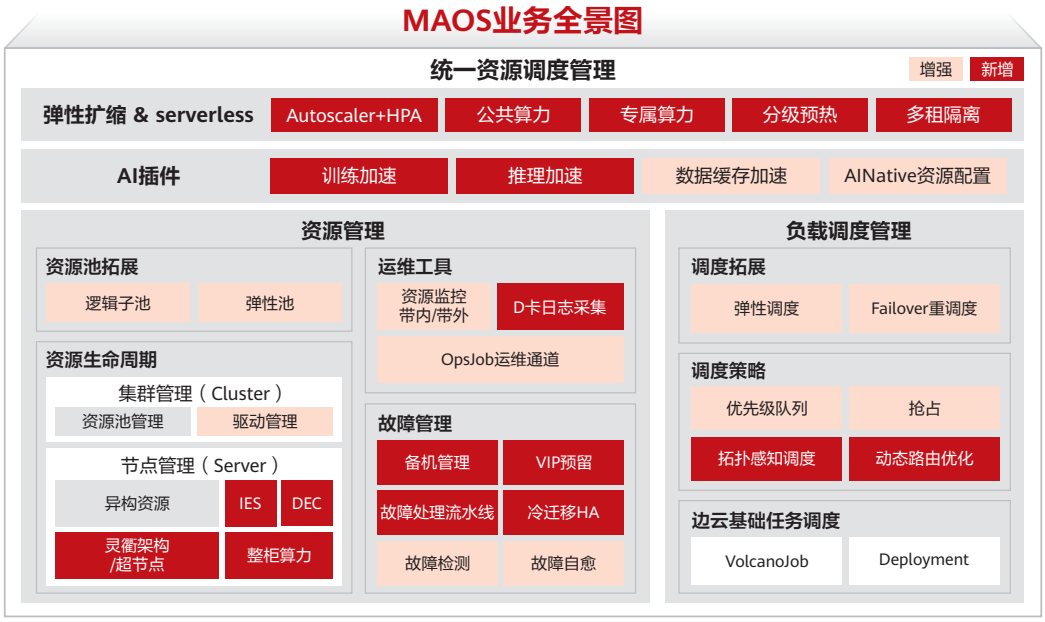


图47 昇腾云算力平台资源管理业务全景图

2) 迁移工具链

迁移工具链是面向昇腾AI开发者提供的工具集，使能开发者高效完成基于昇腾硬件的训练开发、推理开发和算子开发，将之前运行在GPU等硬件平台上的模型迁移到昇腾硬件上，并保持较高的精度和性能表现。

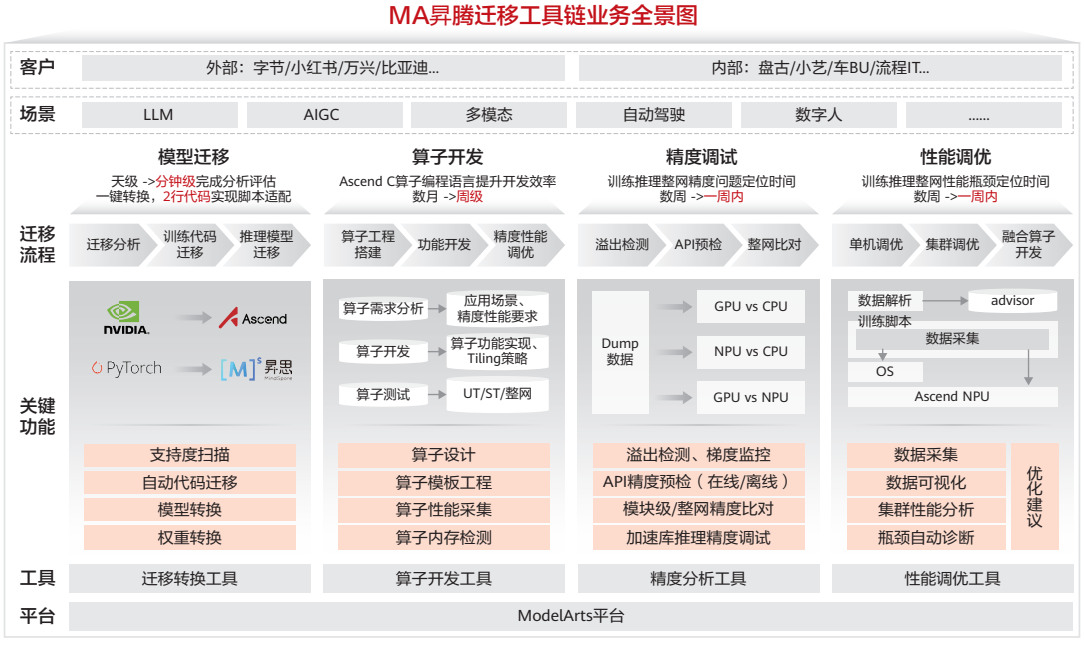


图48 昇腾云算力平台迁移工具链业务全景图

工具链主要包含以下几大类:

- 训练开发工具: 聚焦用户在模型迁移、模型开发中遇到的痛点问题，提供全流程的工具链。通过提供分析迁移工具、精度调试工具和性能调优工具三大主力工具包，帮助用户解决开发过程中迁移困难、精度不达标、性能不达标或劣化等问题，帮助用户解决模型迁移和运行过程中的精度和性能问题。
- 推理开发工具: 提供模型量化、精度调试、性能调优等模型推理开发能力，可快速完成主流推理框架在昇腾平台上的迁移，帮助用户解决推理精度和性能问题，助力用户实现极致推理性能。
- 算子开发工具: 在完备的调试工具和多样的调优数据的帮助下，通过Ascend C的多层接口抽象，简化用户编程难度，助力开发者低成本完成高性能算子开发。算子工具包含算子开发工具和算子编译工具: 算子开发工具为Ascend C编程语言的开发者提供了全面的支持和帮助，使得高性能算子开发变得更加简单和高效。算子编译工具为昇腾CANN提供了算子编译的功能，使得开发者能够将自己的算子代码编译成可在昇腾AI处理器上运行的二进制文件，帮助用户开发自定义算子，获得极致的训练推理性能。

3) 算子优化

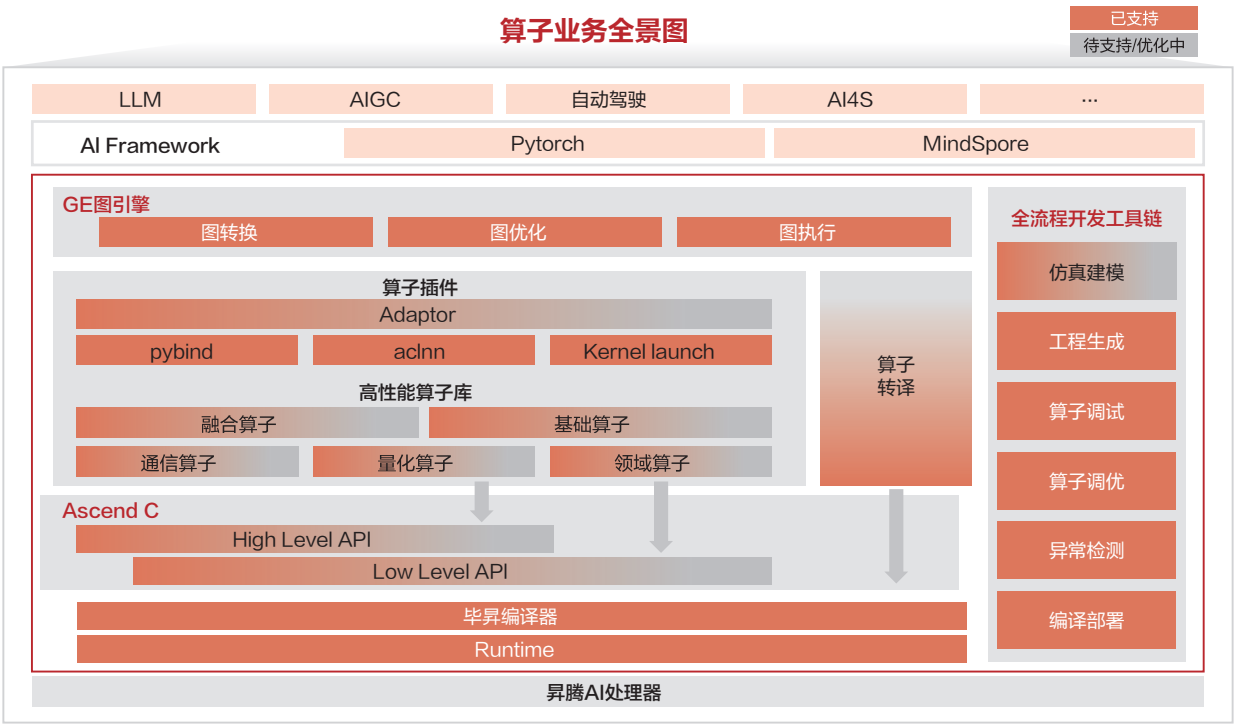


图49 昇腾云算力平台算子业务全景图

i. 规划一：算子覆盖度提升到95%

- 基础算子：gridsampler、deform_conv、mul_sparse_csr
- 融合算子：groupnormsilu、LCCL通信计算融合
- 量化算子：QuantLinear(W8A16、W4A16）

ii. 规划二：极致性能算子库，核心算子加速30%

- Attention融合优化：KVCache管理和FA算子解耦，降低FA算子开发难度，加速Attention模块演进
- MoE-FFN融合优化：合理运用L2 Cache，通过提升L2的命中率，提升SoC的中综合带宽
- 通信-计算融合：通信和计算算子切成更细粒度流水并行
- MSD多尺度反量化：通过MSD技术，解决量化算子因CV分离带来带宽瓶颈问题

iii. 规划三：生态优化，部分算子开发效率加速5倍

- AutoTiling：基于AKG构建Tiling自动搜索调优能力性能
- 编译优化：针对多样化算力、泛化、长尾算子提供极致性能算子代码生成及编译能力
- AscendC API：提供高阶API，减低融合算子开发难度

3.3 AI-Native OS层关键技术解析

AI模型OS层作为AI-Native架构的智能中枢，其核心价值在于构建模型全生命周期的管理体系，使能企业高效实现AI能力的开发与部署。以下将逐层解析这一技术体系如何重构AI开发范式，为产业智能化提供核心支撑。

3.3.1 模型数据处理与准备

数据处理与准备是模型开发的核心环节，随着模型规模扩大、应用场景复杂化以及数据量指数级增长，传统的数据处理模式已无法满足需求，新的技术趋势和核心能力逐渐显现。

» 3.3.1.1 多模态数据融合

随着跨模态任务（如图文生成、视频理解）需求的激增，未来需要更多处理文本、图像、语音、视频等多模态数据的统一表示与联合处理。多模态数据融合是大模型处理复杂现实世界数据的关键技术，它能够整合来自不同模态（如文本、图像、音频、视频等）的信息，提升模型的综合理解能力。

采用多模态数据处理框架，可以将不同模态数据映射到共享嵌入空间（如CLIP），实现不同类型数据的统一表示，通过对比学习对齐图文、音视频数据（如ALIGN），实现跨模态对齐。可以有效解决异构数据融合难题，提升模型泛化能力。

多模态大模型（如GPT-4V、Flamingo、PaLM-E、Kosmos等）依赖于海量、多样化的跨模态数据训练，其数据需求的核心在于规模、多样性和对齐质量。

数据模态类型包括文本、书籍、网页、对话、字幕等（需高质量清洗）；图像/视频，自然图像（如COCO）、视频帧（如YouTube-8M）、医学影像等；音频，语音（LibriSpeech）、环境音（AudioSet）、音乐等；结构化数据，知识图谱、表格、3D点云（如ScanNet）等。

基础模型通常需亿级到万亿级token（文本）和千万级到亿级样本（图像/视频）。例如：PaLI-3使用10B图像-文本对，LLaVA-1.5混合了1.2M图文数据；高质量对齐数据，即使总量大，跨模态对齐数据（如图文配对）需精准标注（如LAION-5B）。

有三个重点考量点：（1）跨模态对齐：数据需明确关联不同模态（如“图像-描述”对、视频-字幕）；（2）多样性：覆盖不同领域（医疗、教育、工业）、语言、文化场景；（3）平衡性：避免模态或主题的偏差（如文本主导而视觉数据不足）。

数据预处理面临三项挑战：（1）清洗：去除噪声、重复、有害内容（如NSFW图像或偏见文本）；（2）标注：自动化（CLIP-style对比学习）或人工标注（成本高）；（3）模态转换，将非文本数据（如音频）转化为模型可理解的嵌入。

未来方向，自监督学习，减少对齐数据依赖（如对比学习）；动态数据混合，根据模型训练阶段调整数据比例；多模态基准测试，构建评估数据集（如MMMU、Seed-Bench）。

» 3.3.1.2 智能数据处理

大模型智能数据处理是指利用大模型（如GPT、BERT、Diffusion Models等）的能力，结合自动化、自适应和智能化的技术手段，对数据进行高效清洗、标注、增强、分析和迭代优化的过程。其核心目标是提升数据质量、降低人工成本、挖掘数据价值，从而支撑大模型训练和应用。

1) 智能数据处理的核心环节

- 数据清洗与去噪：大模型辅助标注：用大模型（如ChatGPT）自动标注或修正错误标签（如文本分类、实体识别）。
- 异常检测：基于嵌入空间距离（如CLIP图像嵌入）或置信度阈值（如LLM生成概率）过滤噪声数据。
- 对抗样本修复：通过对抗训练生成鲁棒性数据（如文本对抗攻击防御）。

2) 数据标注与增强

- 主动学习（Active Learning）：大模型筛选不确定性高的样本交给人标注（如医疗影像分割）。
- 合成数据生成：用扩散模型生成图像数据，或用LLM生成对话数据（如ChatGPT的RLHF阶段）。
- 跨模态对齐：通过多模态大模型（如Flamingo）自动生成图文配对描述。

3) 数据分布优化

- 长尾分布处理：使用大模型对尾部类别过采样（如文本生成补全罕见类别）。
- 领域自适应：通过Prompt调整大模型输出分布（如让GPT生成特定领域文本）。

4) 数据压缩与蒸馏

- 知识蒸馏：用小模型筛选大模型生成的高质量数据（如TinyBERT过滤BERT生成数据）。
- 代表性采样：基于聚类（如K-means on embeddings）选择核心样本。

“以模型能力反哺数据，以数据质量升级模型”。其技术选型需权衡质量、成本、隐私三大维度，实际应用中常采用混合策略。未来方向，构建数据与模型协同进化的闭环系，开发专为大模型优化的数据处理框架，结合区块链存证数据来源，或联邦学习实现隐私保护下的数据协作。随着多模态大模型的发展，未来数据处理将更自动化、自适应和跨模态协同。

» 3.3.1.3 数据回流

大模型数据回流是指在使用大型语言模型(LLM)过程中，将模型输出或用户交互数据重新收集并用于模型改进的过程。这一概念在人工智能领域变得越来越重要，特别是在持续优化模型性能方面。

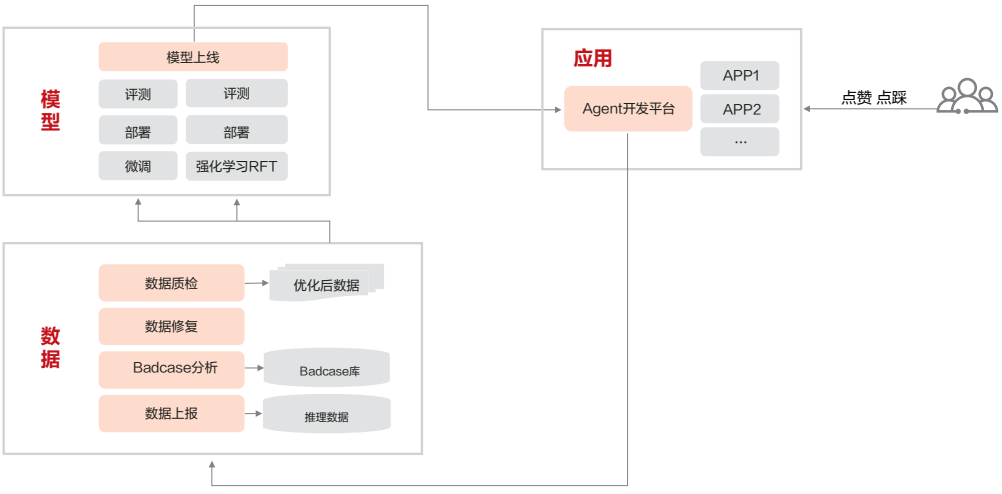


图50 昇腾云算力平台算子业务全景图

数据回流的主要形式包括：（1）用户反馈数据，用户对模型输出的评分、修正或偏好选择；（2）交互日志，用户与模型的完整对话记录；（3）错误案例，模型输出中的错误或不良响应；（4）边缘案例，模型处理困难或不常见的查询。数据回流的作用包括：（1）模型迭代优化，用于微调和改进下一代模型；（2）偏差纠正，识别和减少模型中的偏见；（3）领域适应，使模型更好地适应特定应用场景；（4）安全改进，识别和修复潜在的安全漏洞。数据回流的核心在于构建“数据-模型”协同进化的闭环系统，需在质量、效率、安全间取得平衡。

» 3.3.1.4 数据智能体

新一代的数据应用：数据智能体，核心包括如下几个方面：

- 知识湖：沉淀知识治理GEIR方法论、提供知识治理管理工具。
- 数智大模型：高质量数据集+增量预训练+SFT微调对齐，构建面向数据类任务的专业大模型。
- 数据智能体：研发多轮对话、任务规划、知识检索、内存管理等技术，在各个智能体（数据治理，BI，搜索等）之间共享。

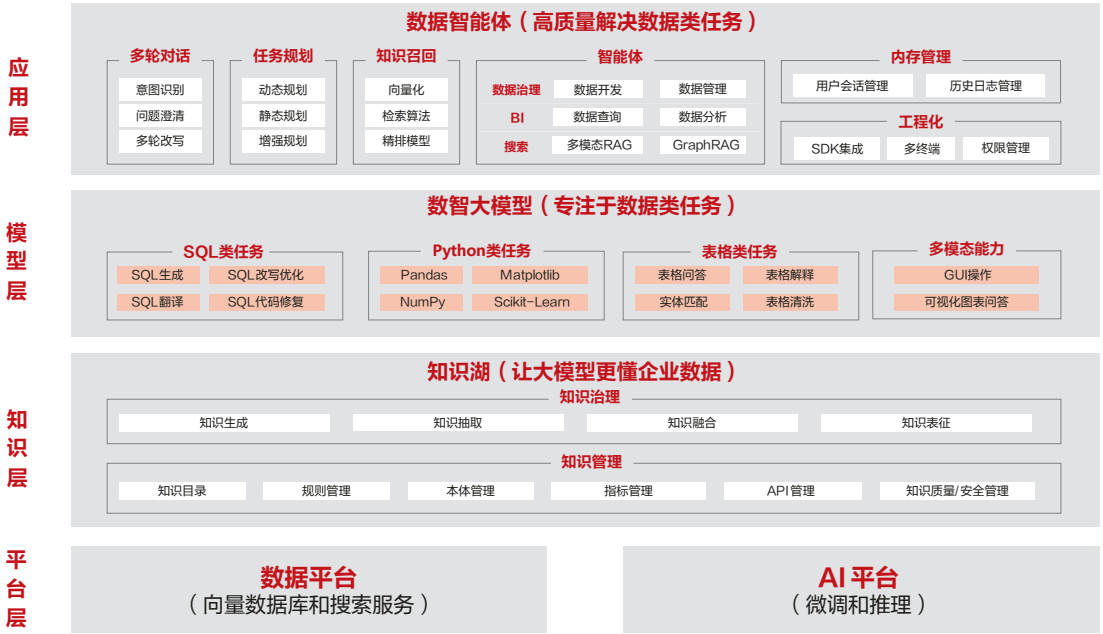


图51 数据智能体

i. 应用层

- 数据生产线全流程：针对数据治理、BI、搜索开发数据智能体，可以处理数据开发、管理、查询、分析、搜索全流程中各类任务，准确率达到90%以上。
- 灵活编排组合：根据用户问题对智能体进行灵活编排组合，利用多个智能体群体的力量处理回答用户的复杂问题。

ii. 模型层

- 高质量数据集：通过抓取、购买、数据合成，构建与数据任务相关的高质量数据集。
- 增量预训练与微调：基于L0模型，通过增量预训练和SFT的方式，来打造面向数据任务的数智大模型，相比基础模型能力提升10%。

iii. 知识层

- 知识治理方法论: 在知识生成、知识抽取、知识融合、知识表征四个方面提炼知识治理方法论。
- 知识治理与管理工具: 利用大模型和工具来辅助知识治理，保证知识治理的效率和效果，效率提升2倍。

iv. 平台层

- 数据平台: 数据平台提供向量数据库、搜索服务，支持数据在不同智能体之间高效流转。
- AI平台: 提供对DataAgent数智大模型的高效微调和推理服务。

1) 构建企业知识湖

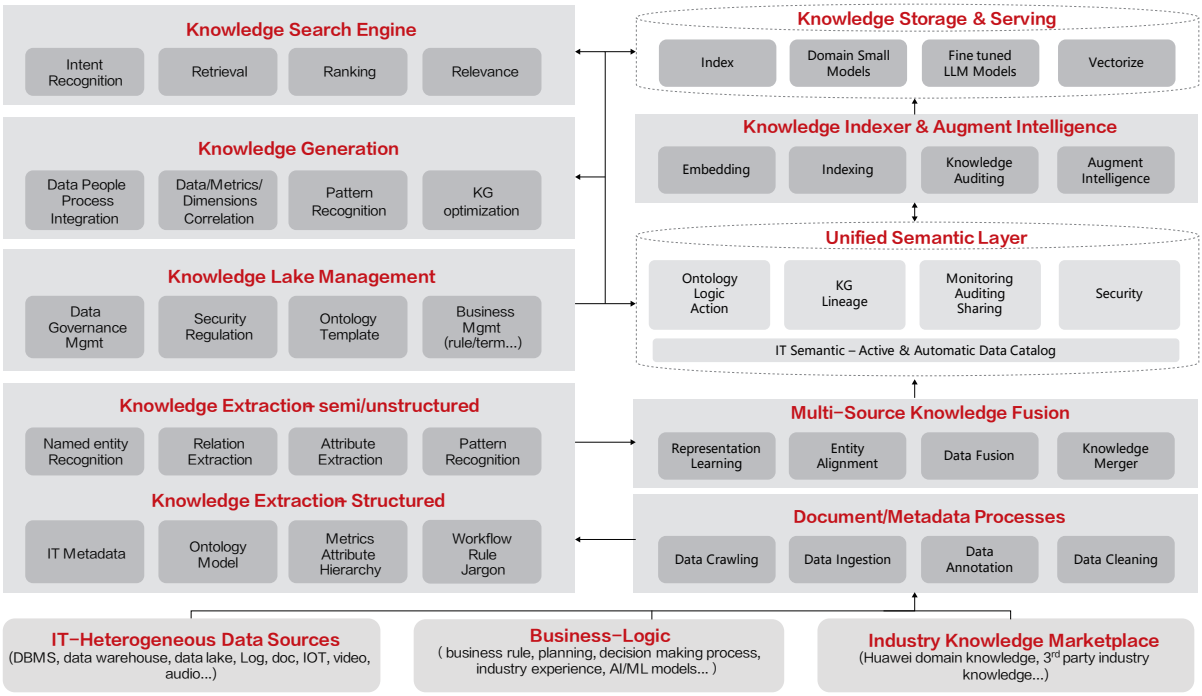


图52 企业知识湖

知识湖智能决策核心技术: LLM + RAG/Graph RAG 使能知识湖构建，基于本体建模构建智能决策流，综合运用向量化和AI Agent能力。

自助式AI数据管理核心技术: 增强数据管理 + 生成式AI 促进自助式数据管理，智能生成数据应用。

关键特性: 业务知识与数据之间的连接器，动态连接任何数据、模型、逻辑、动作

整合和共享工业知识，将BI/LLM/AI/ML与逻辑数据湖集成；数据、指标、维度、业务逻辑和术语的统一语义；统一数据访问-安全控制；搜索和发现、增强智能、知识生成、知识湖管理、监控、审计、共享。

2) 为数智大模型供数

高质量数据集+增量预训练+SFT微调对齐，构建面向数据类任务的专业大模型。

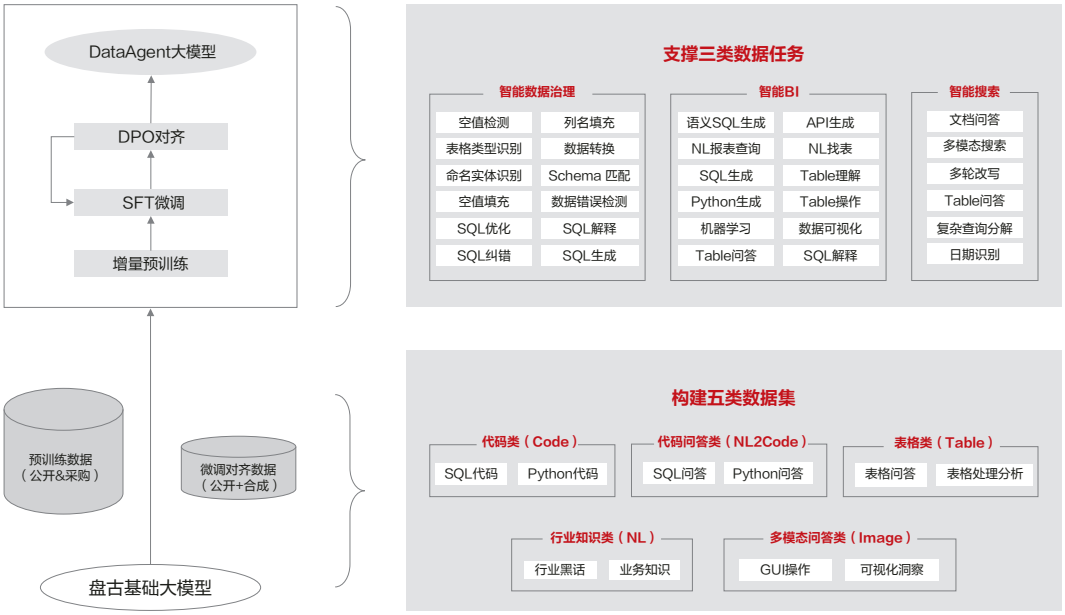


图53 数智大模型

- 五类数据集：代码类(Python和SQL)、代码问答类、表格类、行业知识类、多模态问答类（GUI操作和可视化洞察）。
- 数据采集：采集和购买相关的公开数据并进行高质量的数据清洗。
- 数据合成：通过执行反馈、编程语言翻译、反向翻译等技术自动生成数据。
- 指令进化：通过深度进化、广度进化等技术来增强SFT数据中的指令难度，提升模型遵从复杂指令的能力。
- 基础大模型：在基础大模型的评测中加强数据类任务的评测。
- 预训练数据集：收集五类公开数据集，预训练数据集与基础大模型共享。
- 微调对齐数据：收集公开的微调对齐数据，同时通过数据合成技术来补充更多数据。
- 增量预训练：通过增量预训练，增强模型在代码、表格、行业知识，多模态方面的能力。
- SFT微调/DPO对齐：数据合理配比和选择，通过SFT微调对齐，让模型理解数据类任务的指令，回复满足业务需求。

3.3.2 层次化、可持续迭代的模型训练

» 3.3.2.1 基础大模型

1) NLP大模型

趋势和需求

大语言模型作为当前人工智能领域发展的核心驱动力，近两年迎来了技术和需求爆发式的增长。低成本训练和推理技术的演进带来了算力效率和成本优化的极大提升，通过诸如模型结构、训练策略、硬件亲和算子等优化，模型的训练、推理成本实现了几十倍的降低。但是训练和推理效率还远没有达到理论极限值，未来随着技术进一步的演进，大模型的成本会进一步降低。随着Openai-O1，DeepSeek-R1等推理模型的发布，大语言模型进入了慢思考推理模型的发展阶段，模型通过GRPO/PPO等强化学习训练技术，在模型推理阶段增加长CoT的思考过程，大幅提升数学、代码等相关任务能力。随着技术的逐步演进，大模型会逐步统一、自主融合判断使用慢思考和快思考，从而同时提升泛推理和非推理场景中的能力。

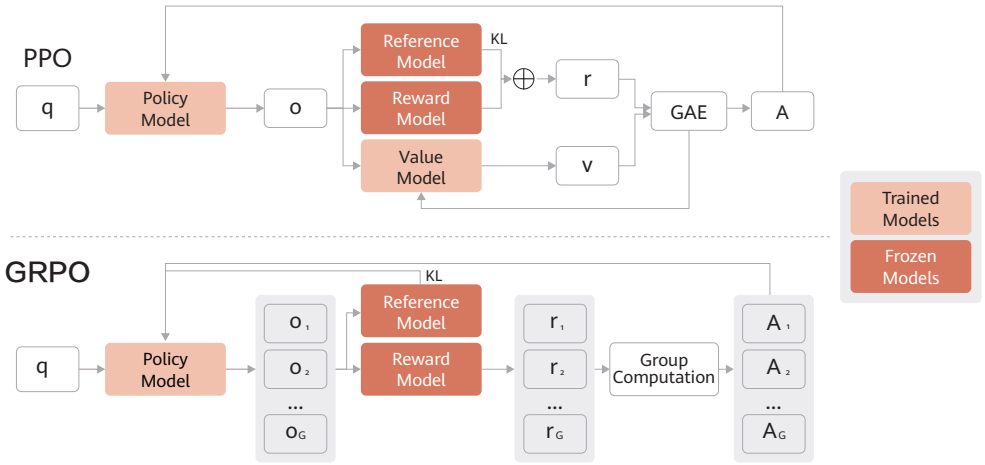


图54 PPO vs GRPO算法示意

随着大模型技术的普及，大模型在金融、医疗等垂直领域的应用越来越精细化、能够处理行业专家任务，在行业大模型构建的过程中，衍生出行业增量训练相关技术，旨在提升行业场景任务能力的同时，尽量减少通用能力的损失。AI Agent是另一个飞速发展的技术方向，大模型通过融合检索、工具调用、记忆功能，通过模型自主规划决策，从而实现例如财报分析、论文综述等更加复杂的专业任务。另外，在更广泛的场景中，只依赖自然语言已经不能满足客户的需求，在此类场景中，自然语言大模型与多模态、具身智能等技术结合的越发深入。

关键特征与相关技术

- 模型结构改进：当前自然语言大模型的模型结构仍然以Transformer-Decoder的结构为主，在此基础上逐步进行各个模块上的改进，包括对attention结构的改进，如GQA、MLA等，来节省推理过程中KV Cache占用显存空间过大的问题。大模型结构上的改进趋势也有一部分向MoE结构发展，使用共享专家+细粒度专家的MoE结构，来扩展模型效果和推理性能的平衡。另外，通过Sparse Attention、Linear Attention等技术提升Attention结构的训练效率也逐步加入模型结构的优化中。

- 预训练：大模型预训练的核心在于通过海量数据+高效架构+自监督学习构建通用表征能力，其技术覆盖分布式计算、优化算法、数据工程等多个层面。分布式计算主要通过并行策略设计、ZeRO等内存优化技术、FP8等低精度训练，进行通信和计算时间掩盖，减少训练显存占用，从而提升大模型训练效率。另一方面，预训练阶段的数据工程技术逐渐成熟，开始更加看重训练数据中的信息量和知识密度的筛选。在预训练过程中逐步加入更多的合成数据，提升训练数据的质量和丰富性。如何使用长CoT数据合成技术，在预训练阶段加入更多的思考过程数据是当前预训练数据研究的重点之一。
- 后训练：后训练主要包括监督微调（SFT）和强化学习两个部分，是大模型实现对话、指令遵从等任务处理能力的关键步骤。监督微调阶段关键部分在于微调训练数据的构建，包含数据合成、数据筛选、和训练数据配比等数据工程技术。强化学习技术近期引起了更加广泛的关注，例如PPO、GRPO等算法，能够提升模型的逻辑推理能力和泛化能力。如何构建稳定高效的大规模在线强化学习训练框架是近期的热门演进方向之一。在线强化学习推动了诸如deepseek-r1，openai-o3等推理模型的发展，通过可控思维长度技术，以及LongToShort技术缩短思考长度，来提升推理模型的推理效率、进行快慢思考模型融合也是当前一个重点问题。
- 行业增量训练：行业增量训练是大模型进行行业落地的关键技术，主要包括行业增量预训练、行业微调、强化微调。对于不同客户的需求以及行业数据储备，对通用模型进行增量训练，提升大模型在行业场景任务上的处理能力，同时尽量保持模型通用能力不下降。在行业增训的过程中，数据筛选技术和动态配比技术能够提升行业增训的效果，同时也能通过精选数据降低训练成本，提升模型训练效率。相对于当前慢思考模型主要针对数学等逻辑推理场景，行业推理模型面临更加复杂的场景和更大的技术挑战。通过行业数据合成和筛选技术构造高质量的可验证的训练数据，通过行业奖励模型、行业数据课程技术进行强化学习训练，都是大模型在行业落地应用中的关键技术。
- AI Agent：大模型通过自我规划、工具集成、多模态融合、记忆与知识管理等技术的融合，完成更加高阶复杂的任务，突破单一模型的能力边界。当前大模型的自主规划、Multi-Agent协同等关键技术还在逐步演进。

2) 多模态大模型（理解）

趋势和需求

随着大模型研究的持续深入，逐步从单一模态向支持多种模态演进。通过在大语言模型基础上扩展对更多模态的支持，多模态大语言模型可以有效融合各种模态信息的感知能力和大语言模型的认知能力，实现更为丰富的能力涌现。多模态大语言模型可以接收包括文本、音频、图像、视频、3D等在内的各类信息，并进行综合理解和处理，更加贴合现实应用诉求，已经成为大模型发展的既定趋势。

在近两年内，随着业界对多模态大模型关注度的持续提升，这一领域取得了令人瞩目的进展。多模态大模型的研究经历了多个阶段，其中包括数据规模的不断扩展，逐步从学术级别的训练数据向工业级别的数据转变；在分辨率方面，从初期的固定低分辨率、固定高分辨率到如今向动态分辨率和原生分辨率的转变，旨在提升模型在处理多模态数据时的表现；训练策略方面，从早期基于LoRA的轻量化训练，逐步过渡到了分阶段的参数更新策略，更精细化地适配模态对齐、多模态融合训练、多模态指令微调等阶段的训练目标，以及多阶段课程学习等方式，通过调整序列长度、分辨率、任务难度等多维度参数来优化模型的性能，同时结合退火等策略进一步提升训练效果。

目前，主流大模型厂商已陆续发布了各自的多模态大模型，整体仍以图文视频模态为主导，3D模态正随着空间智能领域的发展得到更多关注，其它模态仍处于研究探索阶段，尚未融入到主线模型中。整体上，多模态大模型的技术成熟度依然有限，以MMMU-Pro榜单为例，虽然多模态大模型的性能已经取得了显著进展，但与人类专家的表现相比，依然存在一定差距，在一些新型榜单上表现的更加突出，多模态模型在智能表现方面，仍然以继承语言的能力为主，跨模态智能仍处于发展的早期阶段。

这种技术差距一定程度上促进了专注于特定领域的多模态大模型的发展，涵盖了文档处理、工业机器视觉、医疗影像分析、遥感数据处理、具身智能和自动驾驶等多个行业领域。聚焦这些特定领域，能够更加迅速地实现可落地的应用突破。例如，在智能文档和工业机器视觉领域，已有若干应用案例展示了良好的效果，证明了多模态大模型在专业领域中的实际潜力。随着这些领域的不断深化，未来的多模态大模型有望在更多行业中找到更加成熟和可行的应用路径。

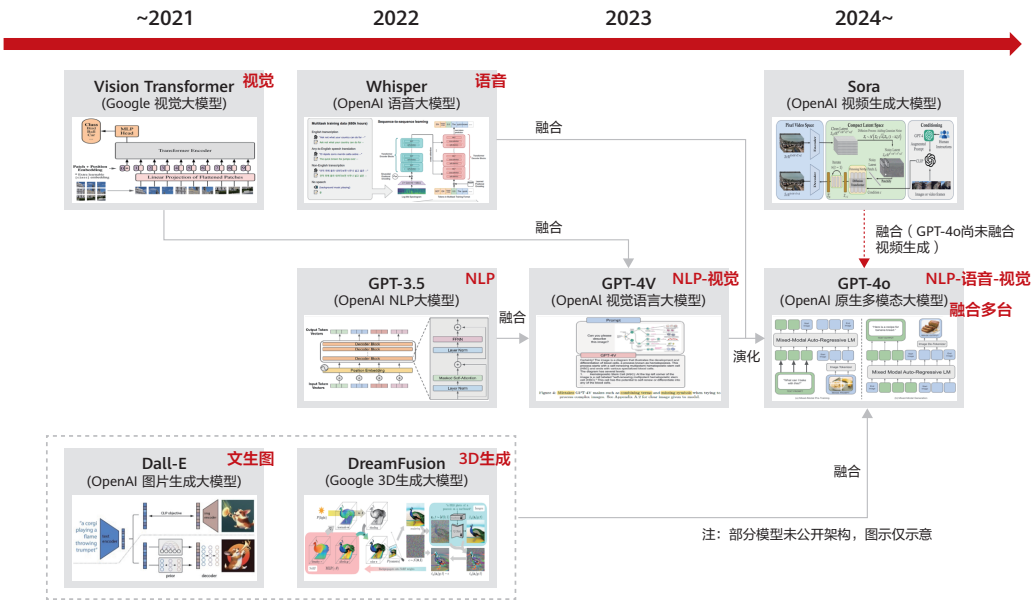


图55 大模型架构发展趋势：从单一模态逐渐探索“多模态大一统”

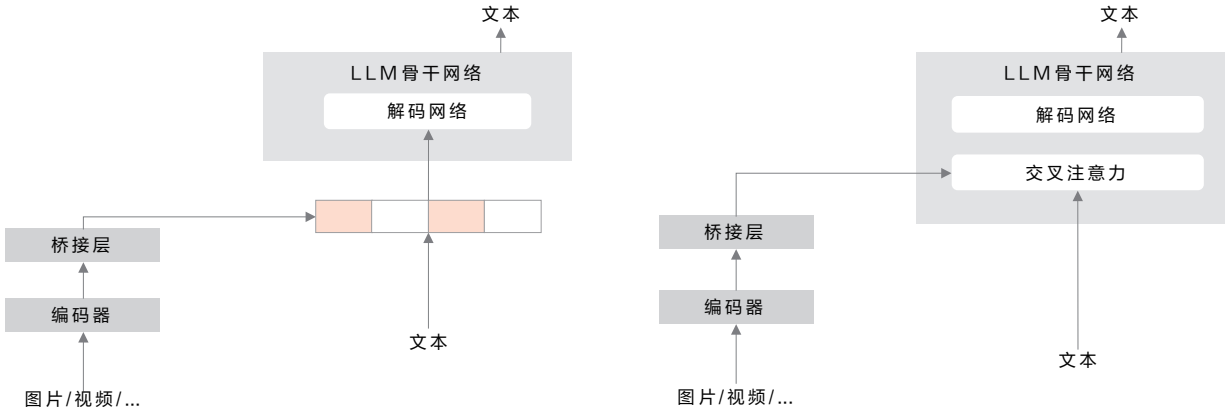
关键特征与相关技术

多模态大模型首要面临的挑战是如何弥合不同模态间的语义鸿沟，其次随着模态数量的增加，多模态任务组合空间呈指数级增长，如何实现充分训练，如何调和不同模态间的相互影响，甚至如何评估模型效果都面临很多新型挑战。

i. 模型架构

多模态大模型通常由编码器、桥接层、LLM骨干网络三部分组成。其关键设计在于如何通过编码器和桥接层完成各模态表征的提取，以及如何将提取后的表征有效融合到骨干网络中。

- 高效模态表征：针对图像和视频模态，业界在多尺度特征融合、原生分辨率、空时位置编码等方面进行了深入的技术探索，实现了高保真的视觉特征提取，极大地增强了模型在时间和空间尺度的感知能力。但是在表征效率方面，仍然面临不小的挑战，特别是在涉及多图、长文档、长视频以及多种模态混合的长上下文场景。通过压缩去冗余，可以提高训推计算效率，提升模型的准确性和泛化性，更重要的是拓宽了模型承载的信息输入范围，从而应对更加综合的多模态应用。对于3D等新模态的引入，如何提取亲和于LLM输入语义空间的表征，需要持续深入探索。
- 模态融合架构：根据非文本模态融入LLM骨干的方式，可以将多模态大模型分为早融合和深度融合两大类。深度融合通过扩展交叉注意力层将其它模态特征注入到LLM骨干。早融合直接将各模态特征与语言特征拼接后输入到LLM网络。早融合可以灵活扩充更多模态，只需要引入模态编码器和桥接层即可；相比之下，深度融合需要在大语言模型网络中扩充模态对应的交叉注意力层，随着模态增多，扩展难度会越来越大。深度融合中交叉注意力层带来了明显的参数量增加，为了实现充分训练，需要更多的训练样本。因构建和训练方面都相对简单，早融合得到了最为广泛的应用。尽管，早融合被证明具备良好的效果扩展性，融合方式改进仍然是一个被广泛研究的课题。



（a）早融合：架构简单，新增参数少，易扩展　　（b）深度融合：架构复杂，新增参数多，难扩展

图56 多模态理解大模型模态融合架构

- 多模态MoE：通过稀疏激活机制，混合专家网络可以在保证模型规模和表达能力的前提下，实现更高的训推计算效率。此外，多模态数据的异构性与MoE结构特性之间有天然的契合度，通过多专家网络隐式地适应不同模态，在缓解模态间冲突问题方面具备良好潜力，在各大厂商的基础多模态大模型，以及VLA等专项模型研究中都得到了广泛应用。

ii. 多模态数据

尽管现实世界中更多的信息是以多模态形式出现的，但大多数多模态模型训练中多模态数据的规模仅为纯文本数据的十分之一，甚至更低。主要原因在于多模态训练依赖跨模态标注数据，标注成本和效率对大规模数据扩展形成了极大障碍。除了持续提升数据标注能力外，在多模态训练数据扩展上呈现出了多路径探索的趋势：

- 无标注或模态不完备的数据：通过引入无监督或弱监督训练，可充分利用互联网上存在的图片、视频、图文交织网页等进行训练。
- 数据合成：可以通过规则方法来合成各类结构化图像，典型的如文档、表格、图表等，根据规则元数据可以快速、高质量建立起合成文档图像与文本之间的对齐关系。
- 数据仿真：对于涉及物理或数字世界交互的场景，可以借助于仿真交互来收集数据，也可以从人类演示视频中提取有效的训练样本。

iii. 多模态训练

- 多模态课程学习：课程学习已经成为大模型普遍采用的训练策略，通过合理安排训练顺序，可以极大提升模型学习效率。对于多模态大模型而言，必须充分考虑模态间的收敛性差异、冲突问题、协同问题等，对训练过程进行有针对性的设计。
- 理解与生成统一建模：理解任务侧重于从各模态中提取信息，而生成任务则通过跨模态的信息整合生成新的内容。统一建模可以将这两者结合起来，通过共享表示学习使得不同模态的信息在同一表示空间中进行交互，增强模态间的对齐和融合效果。

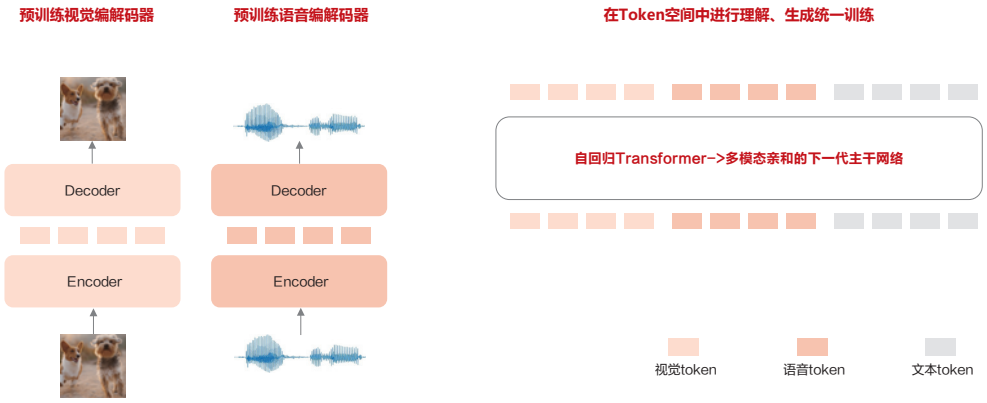


图57 理解与生成统一建模

iv. 高效推理与压缩技术

随着多模态模型规模的增长，其推理效率和部署成本成为关键问题。特别是在文档分析等高吞吐场景，以及生产安全监管等时效性敏感场景。除了大语言模型骨干网络加速外，多模态需要特别考虑编码器网络的推理加速，包括编码器算子优化，编码器分离部署提升吞吐等。

3) 多模态大模型（生成）

趋势和需求

在人工智能蓬勃发展的时代浪潮中，视频和图像生成技术已成为驱动多领域变革的核心力量。早期的图像生成技术受限于硬件性能与算法复杂度，生成的图像质量粗糙，视觉效果不佳且缺乏多样性。随着深度学习技术的兴起，卷积神经网络（CNN）成为推动图像生成领域发展的关键引擎。CNN 通过对海量图像数据的特征学习，能够自动提取图像中的关键信息，显著提升了生成图像的质量，使得图像生成从简单的模式匹配迈向基于特征理解的新阶段。生成对抗网络（GANs）的横空出世更是在图像生成领域掀起了一场革命。GANs 通过生成器与判别器之间的激烈对抗训练，促使生成器不断优化生成策略，极大地提高了生成图像的逼真度，使其能够生成与真实图像难辨真伪的作品，为图像生成技术带来了质的飞跃。近年来，基于Transformer架构在图像和视频生成领域崭露头角，凭借其独特的自注意力机制，能够有效捕捉图像中长距离的依赖关系，让生成的图像在细节和全局结构上表现得更加出色。例如 OpenAI的DALL - E系列模型，能够依据文本描述生成高质量、富有创意的图像，充分展示了Transformer架构在图像生成任务中的巨大潜力。

视频生成技术同样经历了从基础到进阶的蜕变历程。早期的视频生成主要依赖于简单的图像序列拼接或传统的视频编辑算法，生成的视频效果生硬，缺乏自然流畅性。随着Transformer架构的发展，基于Transformer的扩散模型逐渐成为主流。在训练过程中，借助扩散模型对视频数据进行逐步去噪学习，深入理解视频的潜在分布规律。同时通过 Transformer 强大的长序列建模能力，精准把握视频的时空结构并进行生成。这种架构不仅能够生成高分辨率、内容丰富且时空连贯的视频，还在生成效率上有明显提升，为视频生成技术开辟了新的发展方向，代表了当前视频生成领域的前沿探索。

关键特征与相关技术

- 基于自回归生成架构：如下图所示是一种基于自回归的视频生成模型架构，其核心思想是以Transformer为核心的LLM为骨干架构，逐帧或逐token地生成视频内容，每一帧或每一个token的生成都以前面已生成的部分作为条件（文本，图片）。这种机制使得模型能够有效地捕捉视频在时间上的依赖关系，生成具有较好时间连贯性的长视频序列。相比于一些一步到位或固定长度的模型，自回归模型在理论上可以生成任意长度的视频，并且通过条件作用，更容易实现故事线的延续和复杂动态的捕捉。

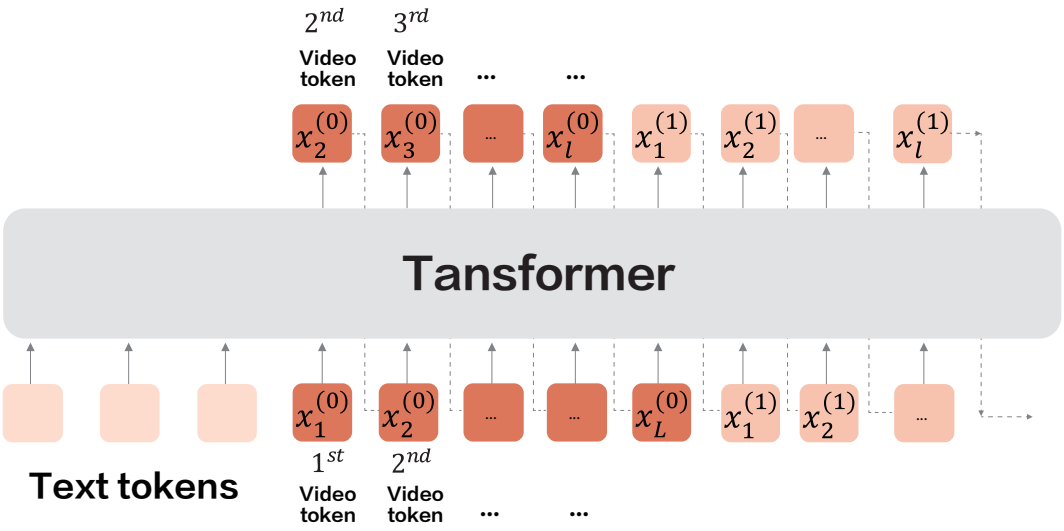


图58 基于自回归架构的视频生成

- 基于扩散的生成架构：基于扩散的生成模型，基本原理是先逐步向数据中添加噪声，然后学习从噪声中恢复原始数据的过程来实现生成。以图像生成为例，首先对一张真实图像逐步添加高斯噪声，随着噪声的不断增加，图像逐渐变得模糊，直至变为完全随机的噪声图像。之后，模型开始学习从这个噪声图像逐步去噪，恢复出原始图像的过程。在生成阶段，模型从纯噪声开始，通过反向去噪过程生成图像。在反向去噪过程中，模型会根据预先学习到的噪声分布规律和去噪策略，逐步调整噪声图像的像素值，使其逐渐向真实图像靠近。这种架构能够生成高质量、逼真且样本多样性好的图像和视频。

如下图所示，一种基于DIT的视频生成架构，主要分为三个部分：（1）视觉空间信息的编解码：原始的视频通过包含时空信息的VAE(变分自编码器)等技术进行压缩和解压。（2）条件注入：通过Re-Captioning技术，对于视频进行更详细的描述，作为条件注入对视频生成进行约束。（3）隐空间加噪去噪：经过压缩后的视觉信息进行Patchify化，加噪之后的隐信息经过DiT(Diffusion Transformer) 进行训练扩展，最后进行降噪。

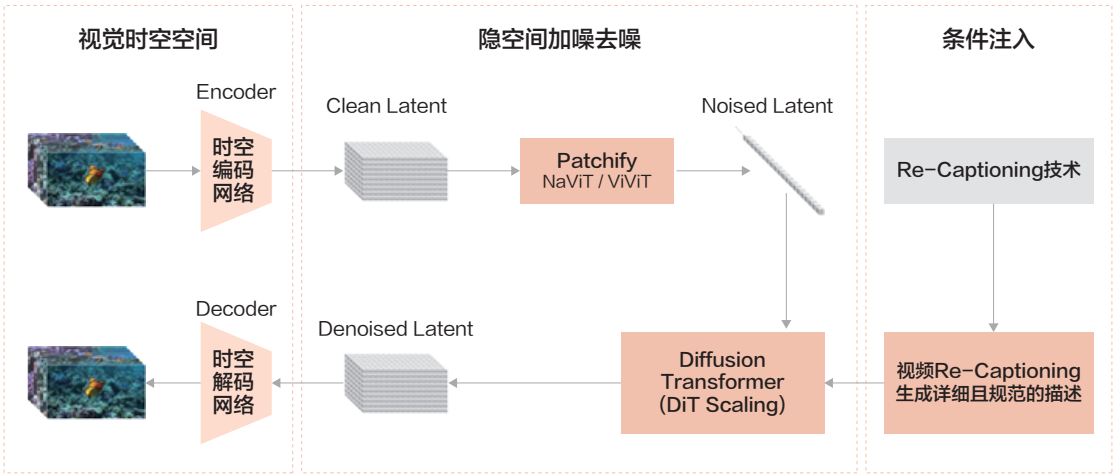


图59 基于扩散架构的视频生成

4) CV大模型

趋势和需求

随着CLIP等视觉-语言模型的兴起，计算机视觉（CV）领域正朝着“大模型”方向快速发展。这些大模型通过大规模预训练，不仅在视觉任务中表现出色，还推动了多模态学习的进步。然而，随着应用需求向实时性、低功耗和高效标注转变，模型的轻量化与数据高效性成为了新的挑战。

- 大模型与轻量化需求平衡：CV 大模型的规模和计算需求不断提升，但实时性与低功耗需求在边缘设备上变得尤为关键。通过模型压缩和知识蒸馏等技术，可以实现更高效的推理，满足低功耗的实际应用场景。
- 长尾问题与数据合成：在许多视觉任务中，数据稀缺导致的长尾问题仍然是瓶颈。扩散模型（Diffusion Models）能够有效合成数据，补充稀缺的样本，从而提升模型的泛化能力。
- 跨任务预训练：借助 CLIP 等多模态模型，CV 领域也逐渐向跨任务和跨模态的预训练方法发展，实现了多任务学习与迁移学习的提升，实现视觉感知任务模型的one for all。

关键特征与相关技术

- 视觉预训练领域经历了从无监督对比学习（如 SimCLR、MoCo）到掩码自编码器（MAE），并最终发展到跨模态对齐（CLIP）的多模态学习范式。对比学习（Contrastive Learning）通过最大化同一图像不同视角之间的一致性，以及最小化不同图像视角之间的一致性，学习鲁棒的视觉表征。这种方法无需标注，仅利用图像自身的信息进行学习。掩码自编码器（Masked Autoencoders - MAE）则通过掩盖图像的部分区域，并训练模型重建这些被掩盖的区域，学习图像的上下文信息和语义表征。这种自监督方法进一步提升了模型对图像结构的理解能力。跨模态对齐（Cross-modal Alignment - CLIP）作为多模态学习的代表，通过对比学习的方式，将图像和文本在同一语义空间中对齐。模型学习哪些文本描述与给定的图像相关联，从而具备了强大的零样本迁移能力，可以直接应用于各种下游视觉任务，而无需针对特定任务进行微调。从视觉模型架构演进来讲，线性复杂度架构提升高分辨率处理效率，Transformer擅长全局建模与多模态对齐。通过扩展模型规模提升性能，并通过设计高效轻量化模型，以支持大模型在在资源受限环境实现高性能实时推理。
- 开集能力的提升，当前的视觉模型主要在封闭集数据上进行训练，这意味着它们只能对已知类别做出预测。然而，在实际应用中，模型往往需要面对未知类别的数据。因此，开集能力成为一个关键问题。为了更好地应对这种挑战，需要设计能够有效识别“未知”类别的模型。这通常包括通过外部数据源增强训练数据，采用开放集识别（Open Set Recognition）技术，或通过自监督学习方法进行预训练，以扩展模型的泛化能力。
- 高效标注与扩散模型在数据生成中的应用，数据标注的瓶颈是当前视觉任务中面临的重要挑战，尤其是当数据样本稀缺时，标注的高成本成为限制模型性能的重要因素。为了解决这一问题，高效标注方法与扩散模型（Diffusion Models）提供了新的思路。通过自监督学习或少量标注数据的增量学习，可以有效减少标注的需求。此外，扩散模型能够通过生成合成数据来补充原始数据集，尤其在长尾分布的场景下，扩散模型能够生成多样化且高质量的图像样本，进一步提升模型的训练效果和数据的多样性。这些方法不仅能够应对数据量不足的问题，还能够解决标签稀缺的问题，从而提升模型的预测能力和鲁棒性。

5) 科学计算大模型

科学研究第五范式，“人工智能驱动的科学研究”（AI for Science），旨在通过人工智能技术的深度融入，推动科学研究的范式变革。

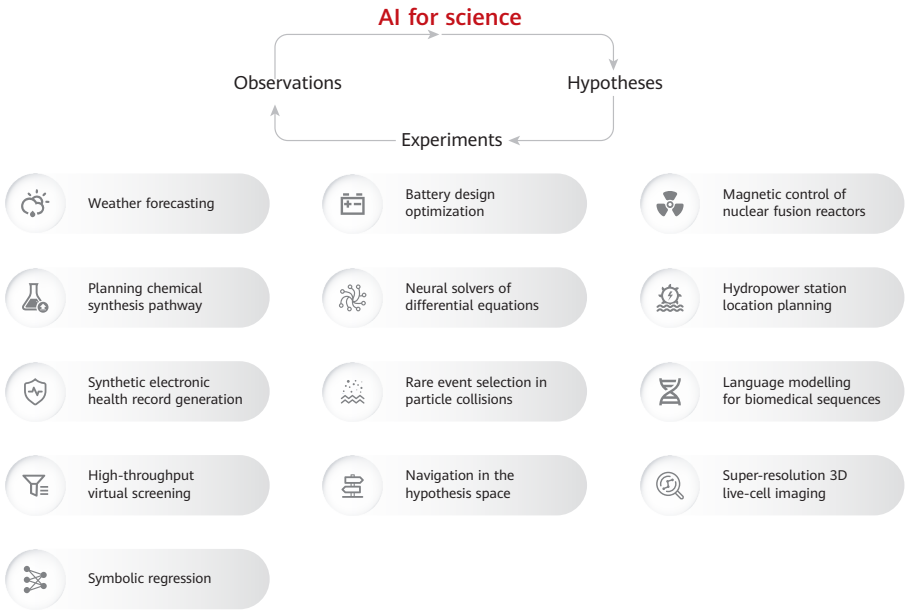


图60 AI在科学研究中的应用

趋势和需求

i. 应对计算复杂性

在许多科学领域，如物理学、化学、生物学等，研究对象往往涉及复杂的系统，其计算复杂性非常高，甚至存在组合爆炸问题。科学研究第五范式通过引入人工智能技术，如深度学习、强化学习等，能够更有效地应对这些计算挑战。

ii. 解决非确定性问题

在科学研究中，许多问题是非确定性的，即其结果受到多种因素的影响，且这些因素之间的关系可能是复杂的、非线性的。科学研究第五范式利用概率统计模型等人工智能技术，能够更准确地描述和预测这些非确定性问题的行为。

iii. 跨学科合作与融合

随着科学研究的不断深入，跨学科合作已成为主流科研方式。科学研究第五范式通过人工智能技术，能够实现不同学科之间的数据共享、模型融合和知识交叉，从而推动跨学科研究的深入发展。

iv. 提升科研效率与准确性

在传统科研范式中，科研人员需要花费大量时间和精力进行数据收集、处理和模型构建。科学研究第五范式通过自动化和智能化的手段，能够显著提高科研效率，同时保持或提升科研结果的准确性。

关键特征与相关技术

i. 自动化实验平台

随着具身智能技术的成熟，自动化的干湿闭环科研将更广泛地应用于生命科学、材料科学等领域，成为科学创新的核心引擎，推动科学发现从“人工驱动”向“自主智能”跃迁。自动化实验平台旨在结合干实验（计算模拟）与湿实验（实际验证）形成闭环，从AI模型提出假设，到自动化实验验证，再到反馈数据优化模型。其中，AI Agent将成为整个科研流程的“超级协调员”，基于大语言模型的推理能力，可自主分解复杂科学问题、规划实验步骤，并调用工具链执行任务；

科学计算领域大模型则是“领域专家”为工具链中的多个环节提供关键能力，包括但不限于正向预测、反向设计等；具身智能则是“手”和“眼”，不仅能够像人类一样再真实的环境中完成精细化的操作，同时能够根据湿实验的状态、做出反应和调整，甚至通过触觉判断产物的性质。进而形成“感知-决策-行动”的闭环能力。

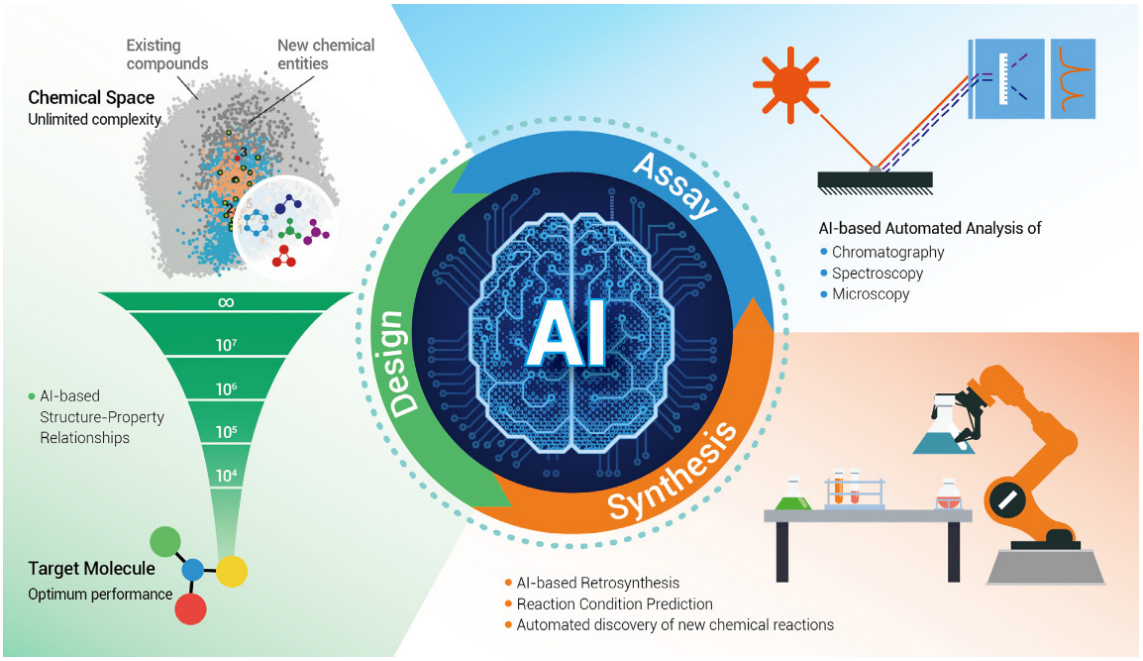


图61 基于AI的干湿闭环自动化化学实验平台

ii. 领域多模态数据融合分析

在科学计算领域，数据的模态具有多样性。如科研报告、小分子结构图、蛋白质结构、材料晶体结构以及气象雷达、卫星数据等。AI处理多样模态的科学计算数据的能力至关重要。不同数据模态之间存在互补的信息。通过结合多种数据模态，可以利用它们之间的互补性，填补各自的不足，从而提高数据分析结果的准确性。

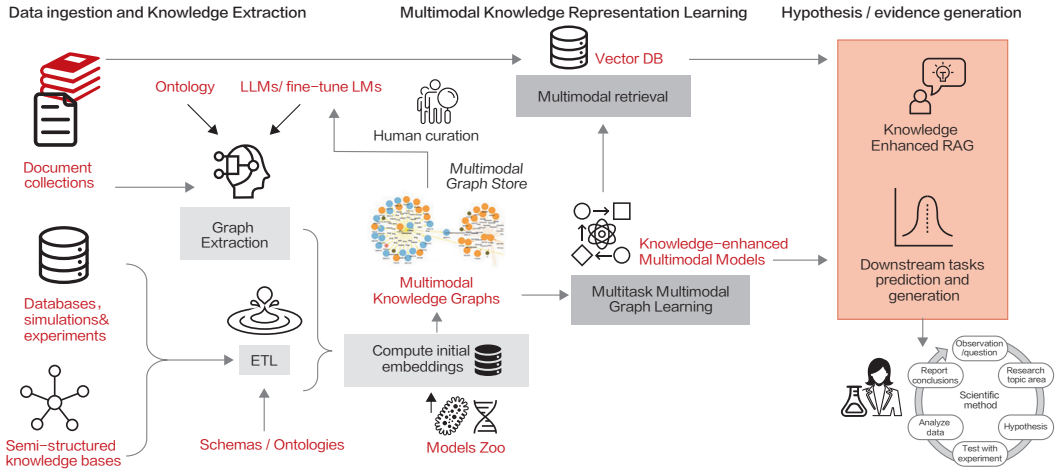


图62 科学领域的多模态数据分析

iii. 符号计算技术与AI融合

符号计算技术用于对符号表达式进行运算，如代数表达式的化简、因式分解、解方程等。这种技术在传统数学研究、物理建模、工程分析等领域具有重要作用。符号计算与AI技术的结合可以增强系统的逻辑推理能力。AI可以提供大量的数据和模式识别能力，而符号计算技术则可以利用这些数据进行逻辑推理和推断，从而得出更准确的结论。

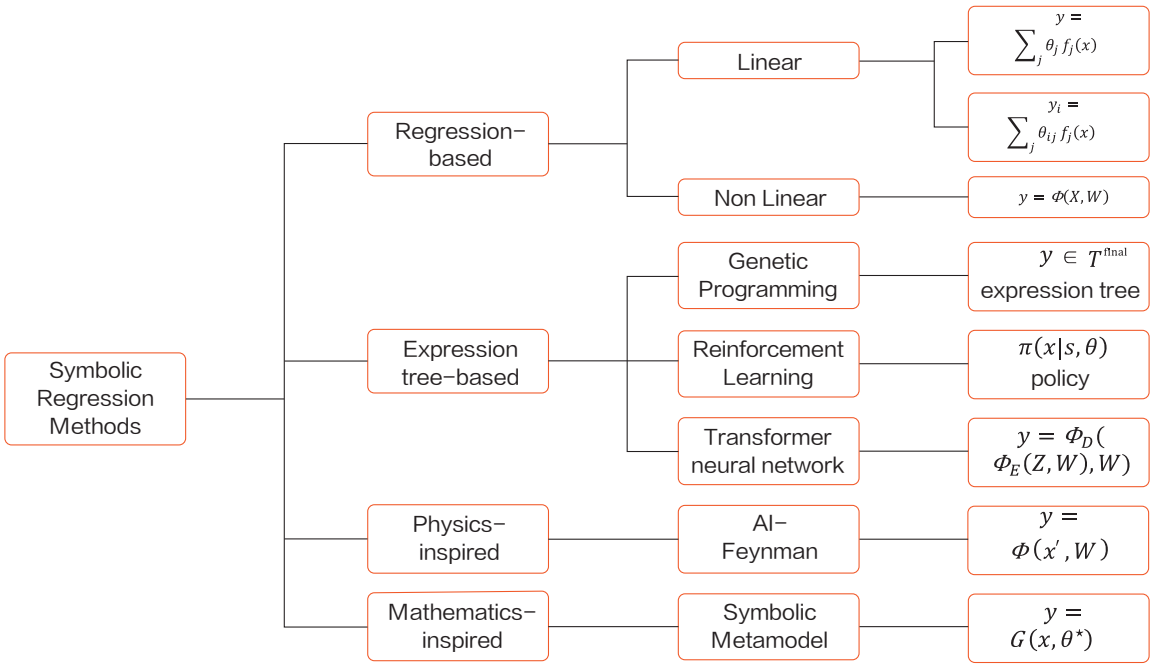


图63 符号计算技术与AI融合

6) 预测大模型

趋势和需求

表格数据预测和时序预测任务广泛存在于各行各业，在商业决策和工业应用中发挥着重要作用。例如，大型零售企业可以根据商品的相关属性，店铺属性、客户画像等构建表格数据预测每日商品的销售量。钢铁生产等工业生产过程可以根据传感器的相关数据构建表格预测工业生产过程的相关参数如温度等，提升生产效率降低生产成本。时序预测专注于时间序列数据的趋势捕捉，在动态系统分析中具有不可替代性。典型应用包括电力负荷预测、股票价格分析、气象预报等需要时间依赖建模的场景。零售商通过分析历史销售数据的时间序列特征，可准确预测未来季度的商品需求，优化库存管理。两类技术在实践中常形成互补：制造企业既用时序模型预测设备振动趋势，又结合设备属性表格数据构建故障预警系统；金融机构则融合宏观经济时序指标与客户属性表格，构建多维度的风险评估体系。

预测大模型主要面向结构化数据的场景，涉及分类回归、时序预测和异常检测等领域。结构化数据广泛存在与真实世界，但是由于大量数据涉及企业等团体的核心数据，难以通过开源方式获取，使得结构化数据相比较自然语言和图像的数量级差距较大，阻碍了预训练模型技术的发展。在时序预测领域，因真实世界广泛存在的时间序列和与大语言模型的语言序列预测任务较强的相关性，时序预测领域率先出现了多项开源大模型和一项国外商业竞品大模型。其他领域因其复杂多变的工况、长短不一的特征信息和或多或少的人工经验，暂时还没有出现大模型的相关产品。为了解决结构化数据的相关预测问题，需要构建一种以结构化数据为主的大模型产品，以期通过大规模的预训练，融合各个领域知识，捕捉潜在工况表征，达到更精准的预测结果，实现商业成功。

关键特征和相关技术：

- 大规模结构化数据生成技术，由于结构化数据收集困难和数据质量难以保证与预训练大模型庞大数据需求的固有矛盾，需要大规模结构化数据的生成技术，期望生成大量的不同分布，特征异构的结构化数据。生成的相关数据用于预测大模型的预训练。
- 跨数据集数据编码，与自然语言和图像不同，结构化数据的不同数据集的数据维度、特征类型、特征含义存在巨大差异，导致当前的模型都是针对某一数据集进行建模，难以构建一个统一的预训练模型。所以需要一个跨数据集的数据编码方案以消除不同数据间的维度、拓扑差异，为进行统一的大规模预训练做好准备。
- 跨任务预训练方法，不同任务数据集需要一个统一的跨任务预训练方案，通过跨任务预训练方法逼近多任务数据的统一表征空间，增益下游任务。下图展示了一种统一编码的方案和对应的跨任务预训练方案。

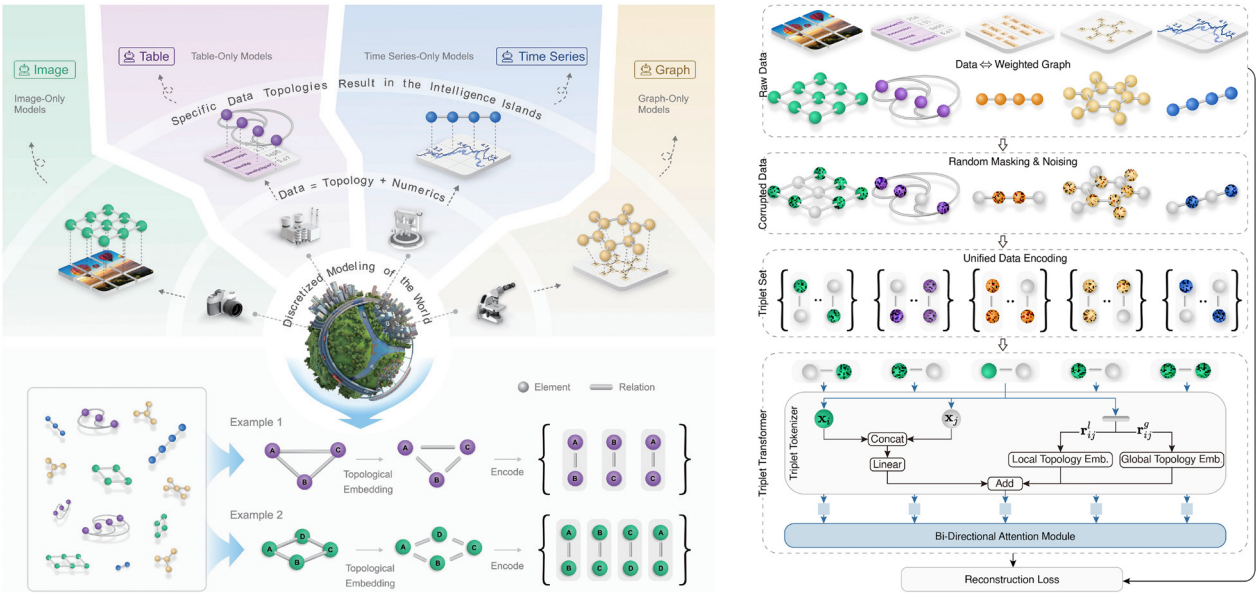


图64 基于三元组编码的统一预训练方案

- 下图是一种基于强化学习的模型微调方法。类似于大语言模型中的偏好对齐方法，该方法是一种对基础模型持续微调的方法，通过生成样本和对应的奖励模型对模型持续微调从而不断提升模型效果，来弥补全量微调、LoRA微调等无法达到的预测残差。

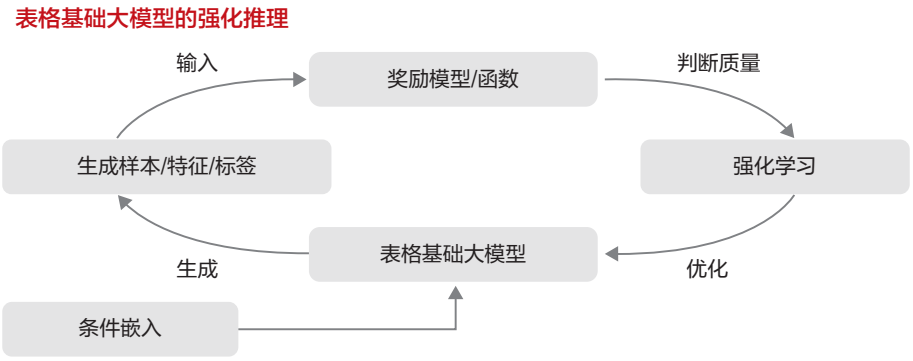


图65 一种基于强化学习的模型微调方法

» 3.3.2.2 AI大模型训练平台

趋势和需求

随着人工智能模型参数规模和训练数据量的爆炸式增长（以GPT-3为例，其参数量已达1750亿），为满足大规模智能模型的高效并行计算需求，并行计算框架经历了持续性优化。当前主流趋势是通过多维度并行技术的协同应用（包括数据并行、模型并行、流水线并行等）实现性能突破。在此背景下，混合并行训练框架（如DeepSpeed、Megatron-LM、Colossal-AI等）应运而生，进一步提升了分布式训练的效率。与此同时，针对长序列输入的处理瓶颈，序列并行技术获得了快速发展。以DeepSpeed的Ulysses和RingAttention方案为代表，这类技术通过在序列长度维度进行拆分并行运算，有效降低了长文本推理的计算复杂度，显著优化了长序列输入的处理效率。

此外，为充分利用节点内的高带宽资源，这些并行训练框架持续优化集合通信技术（如 RingAllReduce、基于树结构的 AllReduce 等）。针对 AllReduce 操作，梯度压缩技术（如 ScaleCom、SketchSGD 等）不断演进，其核心是让压缩后的梯度能直接参与集合通信，从而显著降低通信开销。在任务调度层面，通过通信与计算运算的重叠策略，有效掩盖了一部分通信延迟带来的性能损耗。通信优先级调度技术（如 TicTac 和 ByteScheduler）通过动态调整通信队列的优先级及优化调度算法，以此减少通信延迟。

近年来，随着RL与大规模预训练模型结合的突破性进展，RL训练范式重新成为AI前沿研究热点。该领域主要呈现的技术趋势有：新型RL框架在多模型调度编排和对接多训推后端框架的演进，分离式训推模型部署架构，异步RL流程控制，RL策略优化算法创新GRPO/PPO/...PO，RL环境构建等。

这些新的训练技术的出现对于大模型基础设施提出了多重能力挑战：

- 算力的灵活调度与管理：在实际应用中，需要根据不同的任务优先级和资源需求，灵活地分配和调度算力资源。实现算力的灵活调度，包括计算、网络和存储资源的按需分配和弹性伸缩，提高资源的利用率和训练效率。同时，为达成极致性价比，还需提供serverless训练能力。
- 集群的可靠性管理：可靠性对于大模型训推的稳定性至关重要，需要采用冗余设计、故障切换等技术，提高集群的可靠性。
- 智能运维和监控：大模型基础设施的规模庞大、结构复杂，需要采用智能运维和监控技术来实现对基础设施的实时监控、故障诊断和自动恢复。通过部署监控系统，可以实时监测硬件设备、网络连接、存储系统等的运行状态，及时发现和处理故障。

关键特征和相关技术

训练平台为AI开发者设计，提供便捷的训练作业创建入口，屏蔽底层物理资源运维复杂度。全面兼容多种主流开源框架，实现开箱即用。同时平台支持多种算力资源规格（CPU、GPU、Ascend），配备高性能网络（InfiniBand、RoCE、超节点），并集成高性能文件存储和数据缓存组件。典型的训练平台如下图所示：

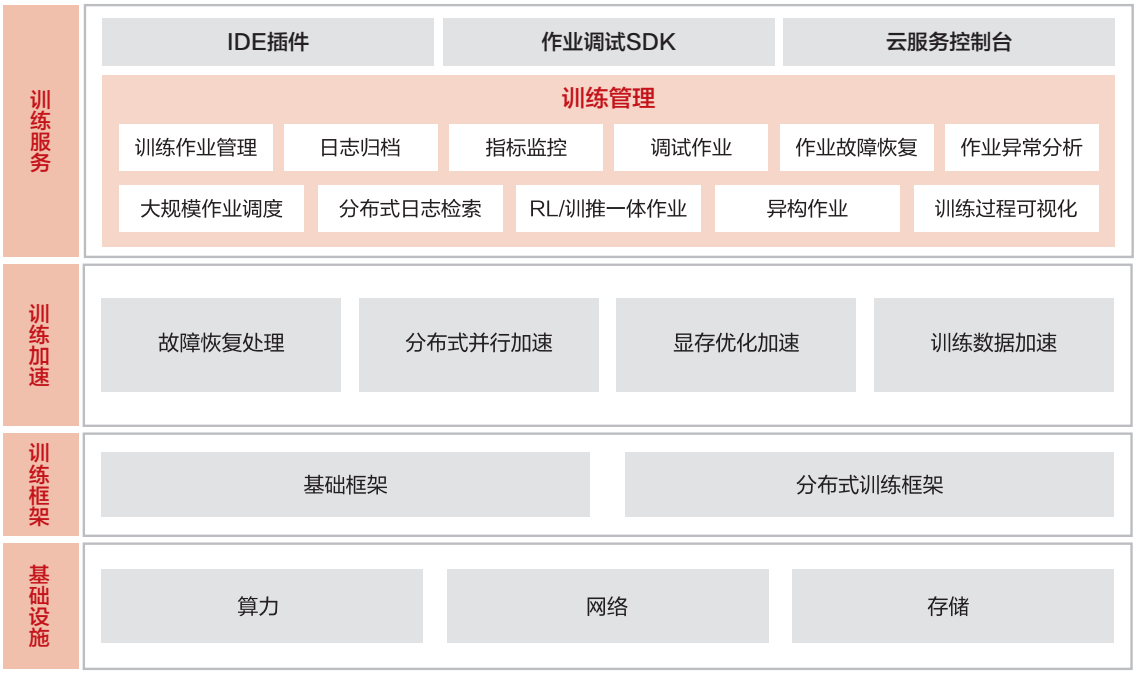


图66 AI大模型训练平台典型架构图

关键能力包括:

- 训练算力调度: 支持AI集群面向训练作业的资源管理和调度，包括计算、网络、存储的弹性分配，支持serverless训练;
- 训练作业管理能力: 支持业界主流AI计算引擎和自研引擎，提供训练作业管理和作业监控以及相关事件日志等信息，支持训练生产模式和调试模式;
- 训练性能加速能力: 支持训练并行策略优化; 支持训练通信无冲突优化; 集成主流性能加速库。
- 训练数据加速能力: 支持训练数据预读取; 训练数据缓存; checkpoint 读写加速; 对接高性能文件系统。
- 训练高可靠能力: 支持运行环境预检与巡检; 巡检发现硬件故障后自动触发训练任务恢复。
- 训练可观测能力: 支持细粒度训练事件记录。支持进程生命周期、堆栈采集观测。支持在线轻量 Profiling 以观测慢卡、慢节点。
- RL 后训练能力: 支持安全沙箱和多样环境，支持主流 RL 框架 verl/Ray/... 及其相关工具链，支持多种 RL 策略更新算法 GRPO/PPO/…，支持 RL 训练加速。

» 3.3.2.3 行业化、场景化大模型的SFT微调及后训练对齐

趋势与需求

大模型在行业领域场景落地的过程中，通用模型往往难以满足实际场景的需求。往往在行业垂域内，客户拥有独有的高质量的数据、有标准的行业方法、行业经验的积累，以及行业独特的、完善的评价体系等等，这些行业特有属性需要高质量、高效率的在大模型应用中进行体现，能够在大模型行业落地中用更小的模型、更低的成本，实现更好的效果。在这种情况下，需要在通用模型的基础上进行增量训练，主要可能包括增量预训练、行业微调和行业强化学习。

典型的需求场景包括：（1）对于需求任务基础模型能力较差，需要在某些Reasoning/Non-Reasoning任务场景进行加强。（2）对于需求任务基础模型已经具备一定能力，需要在特定Reasoning场景进行加强。（3）对于需求任务基础模型能力很差，客户具备大量行业知识和业务过程数据，需要在大部分行业场景进行整体加强。

关键特征与相关技术

1) 行业增训方案

- 典型场景1：通过SFT快速提升大模型行业能力，支持客户利用少量行业数据积累，基于通用或行业推理大模型，针对行业场景快速优化通用模型的行业能力。

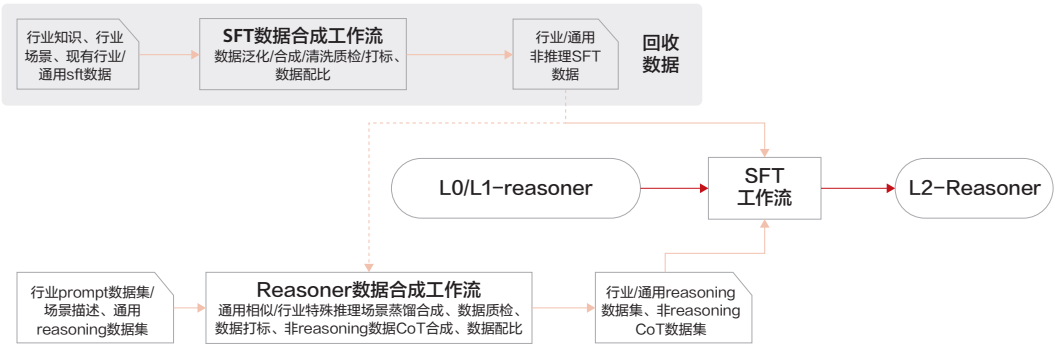


图67 SFT训练流程图

- 典型场景2：通过强化微调RFT深层迭代行业大模型推理能力，支持客户利用少量行业数据积累，基于通用或行业推理大模型，针对特定场景优化推理大模型。

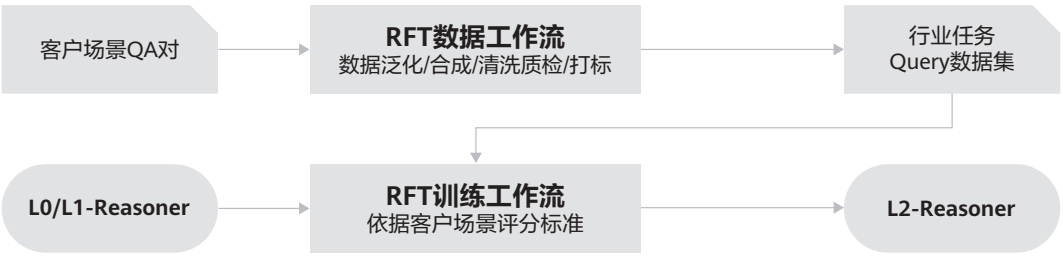


图68 强化微调训练流程图

- 典型场景3: 通过完整的增量训练，全流程提升行业大模型的推理能力，支持客户利用大量行业数据积累，基于通用模型构建行业大模型或行业推理大模型。

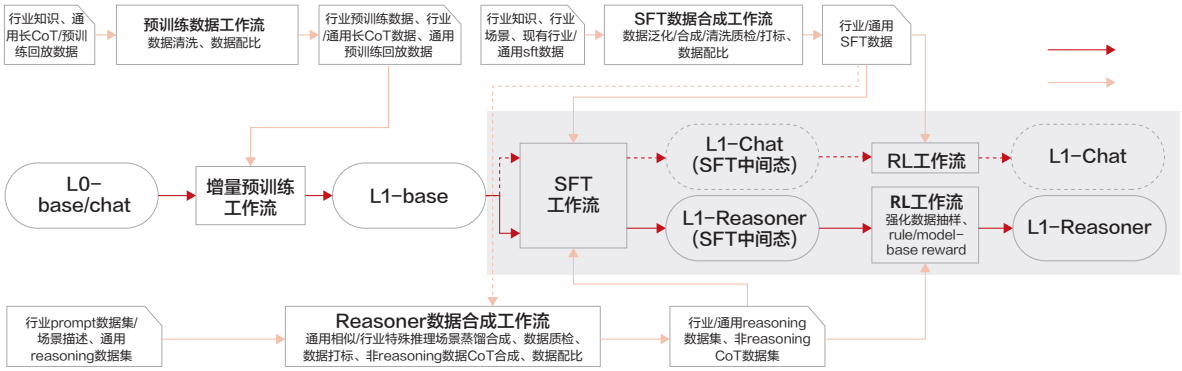


图69 行业增量训练流程图

- 典型场景4: 通过基于人类反馈的强化学习（RLHF）更有效地评估模型质量，从而优化ML模型，使其结果更加准确。

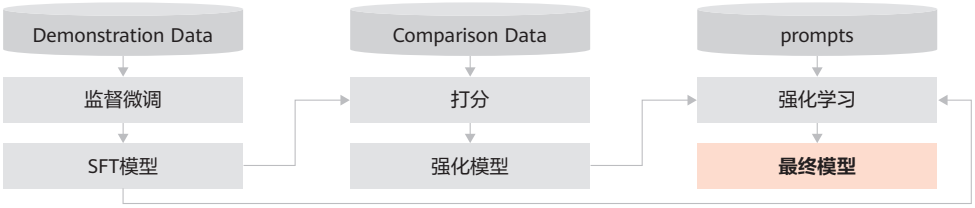


图70 RLFH训练流程图

- 典型场景5: 通过在模型的权重矩阵中添加低秩矩阵(LoRA)来实现微调，从而在不大幅改变原始模型结构的情况下，使其适应特定的任务或数据集。

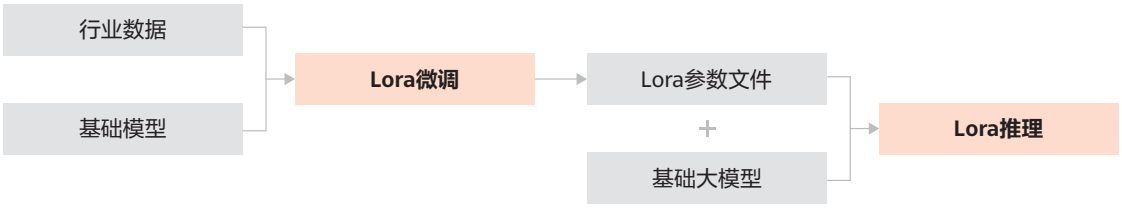


图71 LoRA微调训练流程图

2) 行业增训数据构建

为了满足增量预训练和后训练的需求，需要构建相应的训练数据集。主要数据类型和构建过程如下:

- 增量预训练数据: 用于大模型增量预训练的行业知识数据，需要与通用数据进行配比；数据标准: 行业知识覆盖全面、多样性强、包含行业推理任务相关过程文档。

- 数据构建：（1）准备“XXX 行业”相关的原始文本材料，包括PDF、TXT、XML、Word等格式；（2）如果行业相关文本数据较少，需要通过数据挖掘进行数据扩充，可以通过种子数据从通用数据中采用相似匹配的方式挖掘行业数据；（3）进行数据清洗：采用规则、模型等方式对数据进行清洗、去重、质量评估；（4）数据混合：对数据进行分类、采样后，和通用数据进行混合配比。
- Reasoning SFT数据：混合R1等推理模型合成数据和人工标注数据使用，按照CoT思维链格式，通过Prompt输出推理过程（在<think>标记符内），最终结果（在<answer>标记符内）；数据质量标准：数据答案正确率高，Prompt多样性强，数据覆盖大多数行业任务场景类型。
- 数据构建：（1）Query生成：根据行业场景、SFT种子数据、行业文本素材，通过Self-QA、自然问答挖掘、指令泛化等方式自动生成行业相关Query，或者通过人工标注生成优质行业场景Query；（2）Query筛选：通过分类、打标、聚类等方法，筛选出高质量、多样性强的Query数据集；（3）长CoT数据生成：使用DS-R1等推理模型，生成长思维链的回复数据，回复数据包含<Query, Think, Answer>；（4）数据筛选：从生成的回复中筛选出回答正确的优质数据。对于能够通过规则判断答案的场景，可以通过规则、Rule-based Reward进行判断筛选；对于其他Reasoning场景，可以通过生成式奖励模型和标准答案，或者人工标注的方式，对回复进行打分，筛选出回复正确的数据。
- Non-Reasoning SFT数据：对于Reasoning模型构建中的Non-Reasoning SFT数据，同样需要构建类似的CoT数据（对于非常简单的类似“Hello”的场景，可以不包含CoT过程）。混合合成数据和人工标注数据使用，按照提示调优对格式整理。数据质量标准：数据答案正确率高，Prompt多样性强，数据覆盖大多数行业任务场景类型。
- 数据构建：数据构建的流程与Reasoning SFT数据构建流程基本一致。主要区别：在CoT数据生成过程中可以采用推理模型DS-R1或者通用模型DS-v3通过Prompt工程的方式生成满足要求的CoT数据，在数据筛选阶段，可以通过生成式奖励模型+参考答案进行评分筛选，或者通过人工参与标注的方式提升SFT的数据质量。
- RL数据：RL数据可以复用SFT数据生成的Query和Response。
- RFT数据：对于有明确答案且答案可以使用规则判断的Reasoning场景，可以通过RFT训练进行特定场景任务的效果提升。数据需要少量（如几千条）该任务场景的QA数据，包含Query和对应的标准答案（不用包含CoT数据）。数据质量标准：数据答案正确，Prompt多样性强。

» 3.3.2.4 模型监控、更新与治理框架

趋势与需求

2018 年 BERT 开启了预训练 Transformer 的时代，随后不断涌现出 GPT-2、GPT-3、 PaLM 和 LLama、Qwen、DeepSeek 等模型。OpenAI 公司提出的自然语言模型性能 Scaling Laws 牵引业界持续提升模型参数总量、训练算力以及数据量。以获得更好的模型性能。而随着模型的参数量持续提升，训练、推理复杂度也会增加，模型流行趋势也从稠密模型为主，转变到稀疏 MoE 模型。当前随着 DeepSeek R1 的开源促进后训练技术的普及，后训练的 Scaling Laws 开始显现巨大的模型性能提升空间，基于 RL 的后训练帮助模型的复杂“思维”能力快速提升。

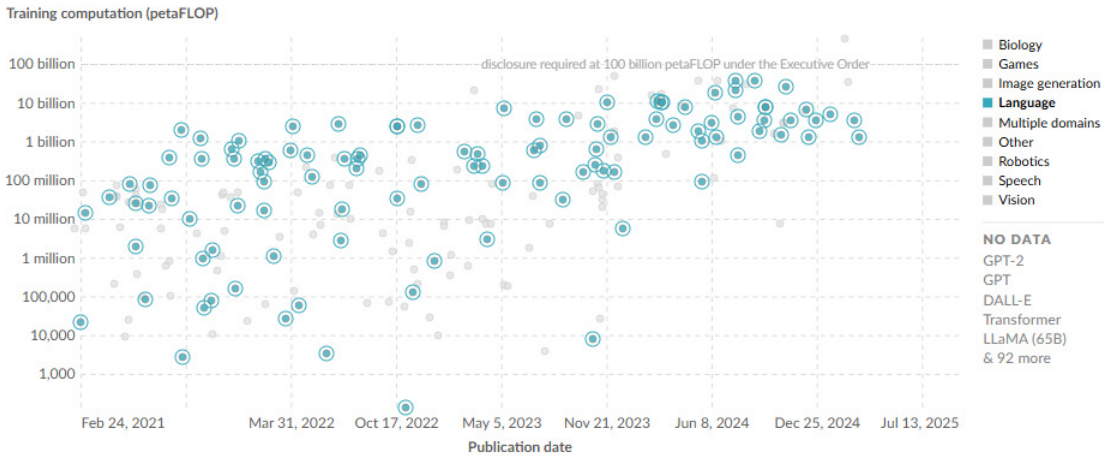


图72 大语言模型参数规模趋势图

近年来，随着模型推理能力持续提升，带来语言模型应用的旺盛需求。

来源：https://www.china-aii.com/yjbg/7140791.jhtml，中国工业互联网研究院

图73 开源模型数量、工具和社区趋势图

关键特征与相关技术

随着模型数量的增加，原先的模型库逐渐演变为了一体化的模型开源社区，全方位为模型应用与模型开发提供一站式支持。模型开源社区提供含算力、AI框架 、模型、数据服务的流程化服务。用户通过社区查看流行模型榜单、模型在不同专业数据集上的测试得分；基于社区提供的算力，一键轻量部署模型进行实际效果评测。用户如需根据自有业务特征数据进行增量训练，还可以下载模型代码与权重；参考模型 README 完成数据处理、本地化的增量训练及推理部署过程。

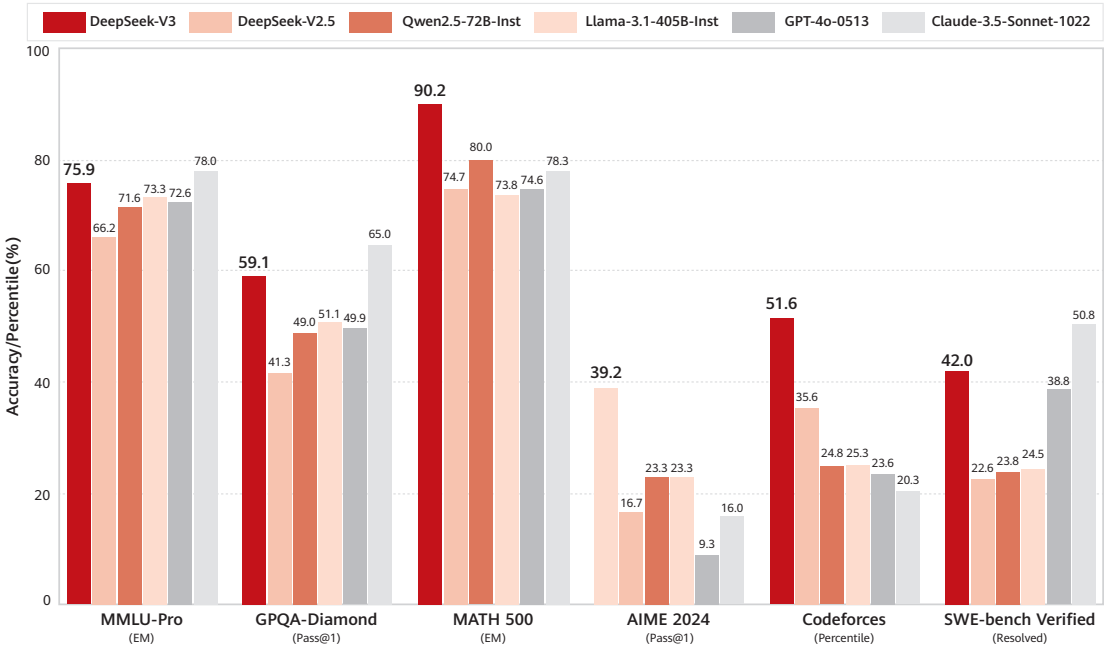


图74 DeepSeekV3 模型在不同专业数据集上的评测得分

当然用户也可将自己的模型发布到模型社区中，与广大的AI开发者一同提升模型的能力。对于优秀的开源社区，开源协作机制是基础，还需要有优质资产、易用工具、可持续的生态机制，才能吸引开发者持续贡献能量。模型社区可提供标准化的模型资产开发框架降低模型开发与发布、共享门槛；通过开源协作，保持模型数量及模型持续的领先性，持续吸引用户使用。

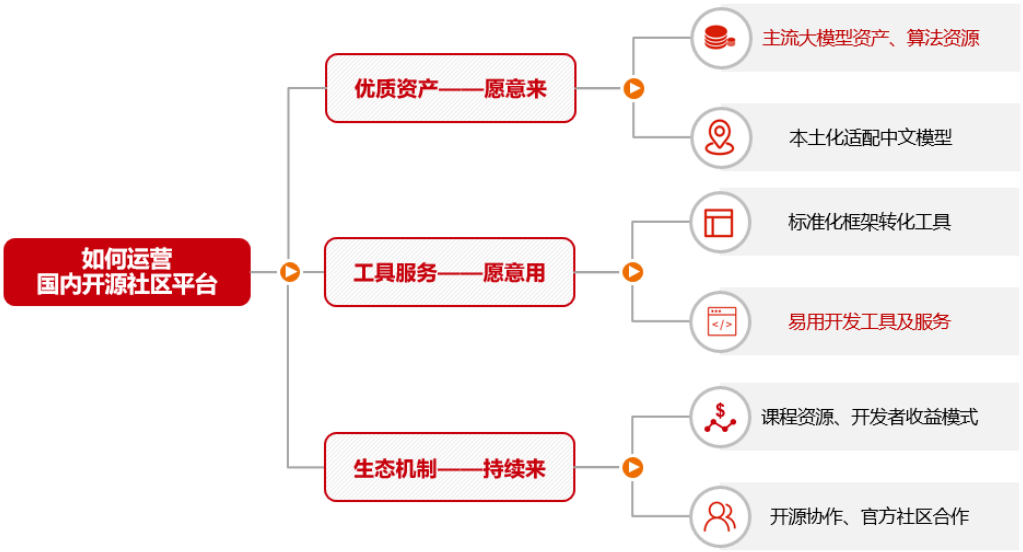
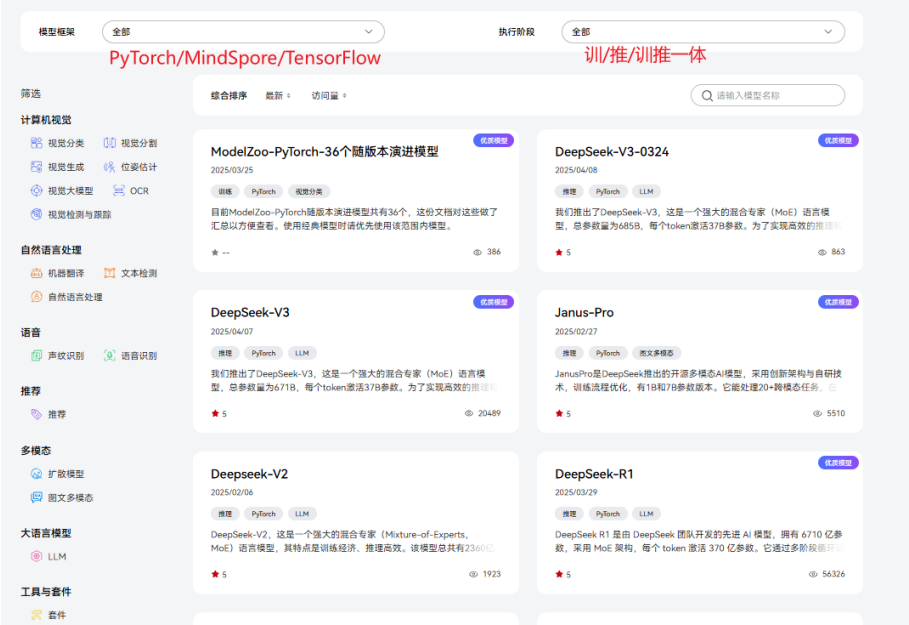


图75 模型开源社区运营方案示意图

业界的模型社区缺少昇腾算力亲和的模型，而昇腾模型社区提供众多亲和昇腾算力的模型资源，覆盖训练、推理、训推一体（后训练强化学习）场景，支持众多开源模型在昇腾算力上的高效增训与推理。



来源：https://www.hiascend.com/software/modelzoo/models

图76 典型模型社区示例图

模型开发需要使用到算力资源以及各类调测工具，用户可基于 AI 平台进行 AI 业务 CT/CI/CD 实践探索，提升模型开发迭代效率：

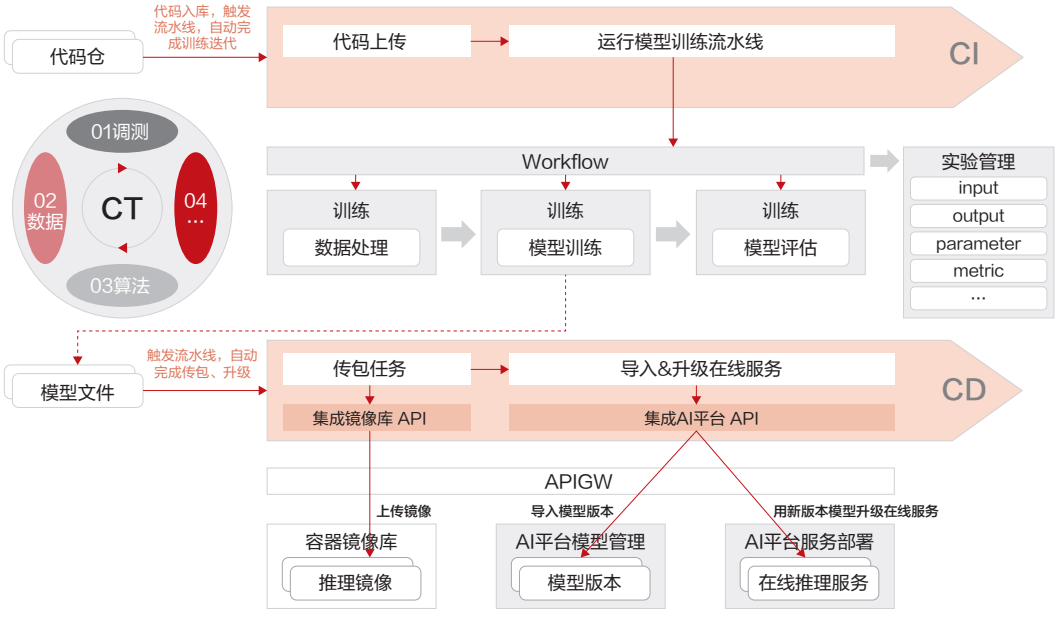


图77 典型模型开发流程图

模型开发完成后，稳定的模型代码和权重通过模型库平台进行共享和复用。典型模型（资产）库平台分层视图：

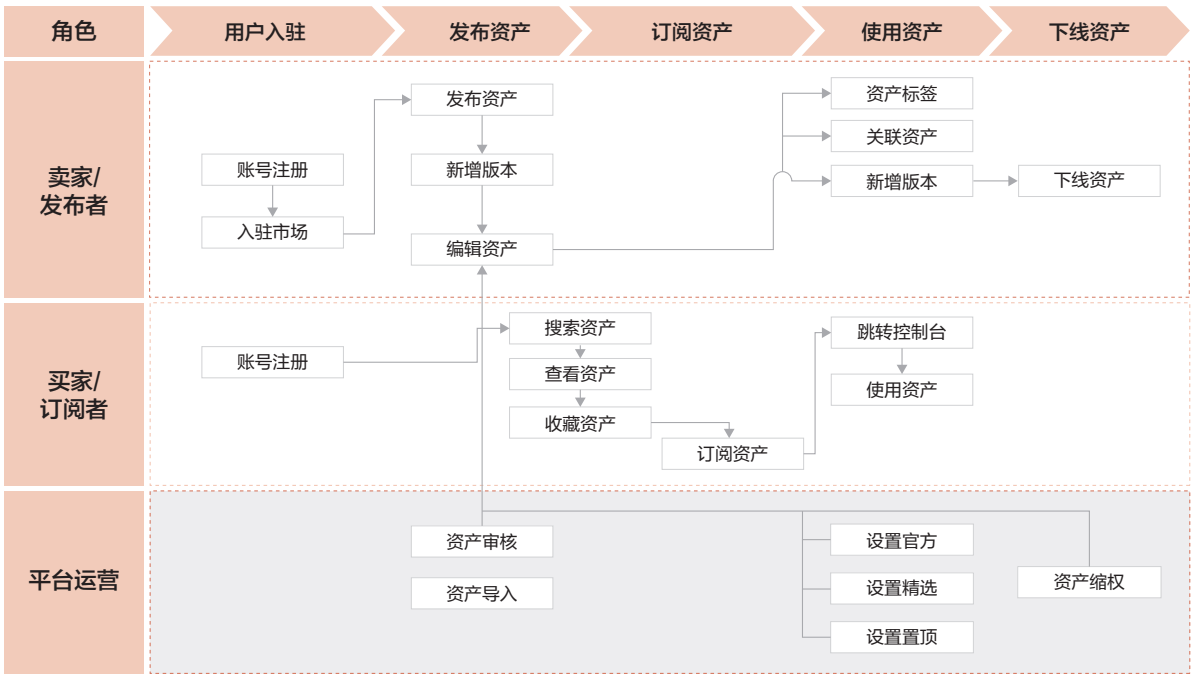


图78 典型模型（资产）库平台分层视图

典型的模型库平台关键能力应该包括以下几个方面：

- 资产发布：面向开发者提供统一的共享平台和开发者友好的资产开发工具；
- 资产管理：提供灵活在线订阅、更新、修改、审批的资产管理功能；
- 资产使用：支持用户模糊搜索资产，查看热门资产及订阅资产开发，在完成订阅后可以一键部署模型评测任务；
- 资产共享：提供资产的权限管理能力，支持发布者设置资产分享的范围。

3.3.3 弹性按需的Serverless化模型推理服务

趋势和需求

当前基于Transformer架构的云上大模型推理系统的难点和痛点是如何在不影响服务性能的同时最小化资源分配。当分配的计算资源数量大于服务的请求量时，不可避免地会造成资源的浪费，当分配的计算资源数量小于服务的请求量时，模型服务的请求不可避免地会堆积，导致服务性能受到影响。然而，由于AI服务请求具有动态性，且一台计算服务器所能处理的请求数量有限，很难准确估算某一时间节点请求所需的计算资源，并提前准备好相对应的资源。一个模型服务请求存在以下两方面的波动性：

- 请求量随时间的波动性：和传统应用一样，用户（客户端）调用推理服务API的时机存在不确定性，这在服务器端表现为推理服务所收到的请求随时间剧烈变化。从内网业务数据看，服务器收到的请求剧烈波动，且波动的间隔非常短：在10秒内，服务收到的请求量会增长5倍。
- 单个请求所需计算资源的波动性：即使服务器端收到的请求数量恒定，不同请求所需的计算资源也不同。例如，在LLM中，处理请求所需的计算资源会随着context的长度变化。同时，由于其auto-regressive的特性，decode阶段产生输出所需的时间和资源也不同。

在传统的非Serverless范式中，系统为每个模型实例分配固定的计算资源，并在整个生命周期内持续运行推理服务进程。这种范式下，为了应对不同业务负载，平台需要事先预估模型热度、预分配资源、维持服务可用性。虽然这种方式简单直观，但在大模型推理场景中暴露出显著的局限性。首先，资源利用率低。大模型往往需要高性能计算资源（如数百GB显存的GPU集群），而长驻进程在低负载或空闲时段仍占用这些资源，导致资源浪费严重。其次，弹性差。模型负载呈现高度动态变化，例如请求突发、冷启动频繁、推理需求跨时间段剧烈波动。非Serverless范式下，模型加载耗时大、服务扩缩容响应慢，难以支撑这类动态负载。最后，模型多样性带来的运维复杂性不断增加。在多租户、多模型场景中，如何在有限资源下同时调度多个模型，避免资源冲突、满足服务质量要求。

这些问题催生了Serverless LLM推理范式的需求。Serverless强调按需加载、即时调度、自动伸缩。在该范式下，模型无需常驻进程，而是根据请求动态加载、执行后释放资源，实现真正的以请求为中心的资源调度方式。通过结合弹性资源池、高效的模型加载与缓存机制、调度器感知的负载感知分发策略，Serverless范式可显著提升资源利用率、降低成本，并提高系统对突发请求的响应能力。此外，它为LLM推理服务的标准化、平台化提供了天然的技术基础，更易于构建统一的推理基础设施。这些优势使得Serverless LLM推理成为未来云端大模型服务的发展趋势。

关键特征与相关技术

本章介绍Serverless AI平台中的推理引擎架构。

1) 推理引擎 (以下简称FS)：

- FS架构分3层，最上层为调度层，中间为计算层，底层为数据层。其中调度层主体以Python为主，维护了请求的调度队列，并且内置了大约3~5种调度算法。中间计算层以前后处理，模型计算为主。数据层负责本地HBM/DRAM分配管理以及中间数据的管理；
- FS采用了全异步化的架构设计，并且采用了数据和控制分离的设计。数据部分由RTC承担；
- Relational Tensor Cache(RTC)：RTC负责推理引擎的数据管理，特别是本地HBM、DRAM的内存分配管理，RTC同时负责推理实例之间的数据传输等。

2) HBM池化：RTC实现了推理实例内和推理实例间的HBM池化、互通。

- RTC支持推理实例内部和推理实例之间的HBM互通，使得计算任务能够灵活地访问共享的高带宽内存资源。推理实例间的互通机制确保了跨任务的数据传输高效而稳定；
- 一套统一的HBM间数据流动抽象可以支撑各种分布式推理技术（例如经典的PD分离，D到P的回传，Sequence Parallelism均可基于RTC搭建）；
- HBM池化的核心优势：资源复用率提升，通过池化机制，多个推理任务可以动态分配和复用HBM资源，减少了内存分配的碎片化问题。内存利用效率优化，相比传统的静态分配模式，池化大幅降低了资源闲置率。灵活性增强，池化支持按需动态扩展资源分配，能够适应不同模型推理任务对内存大小和带宽的需求。

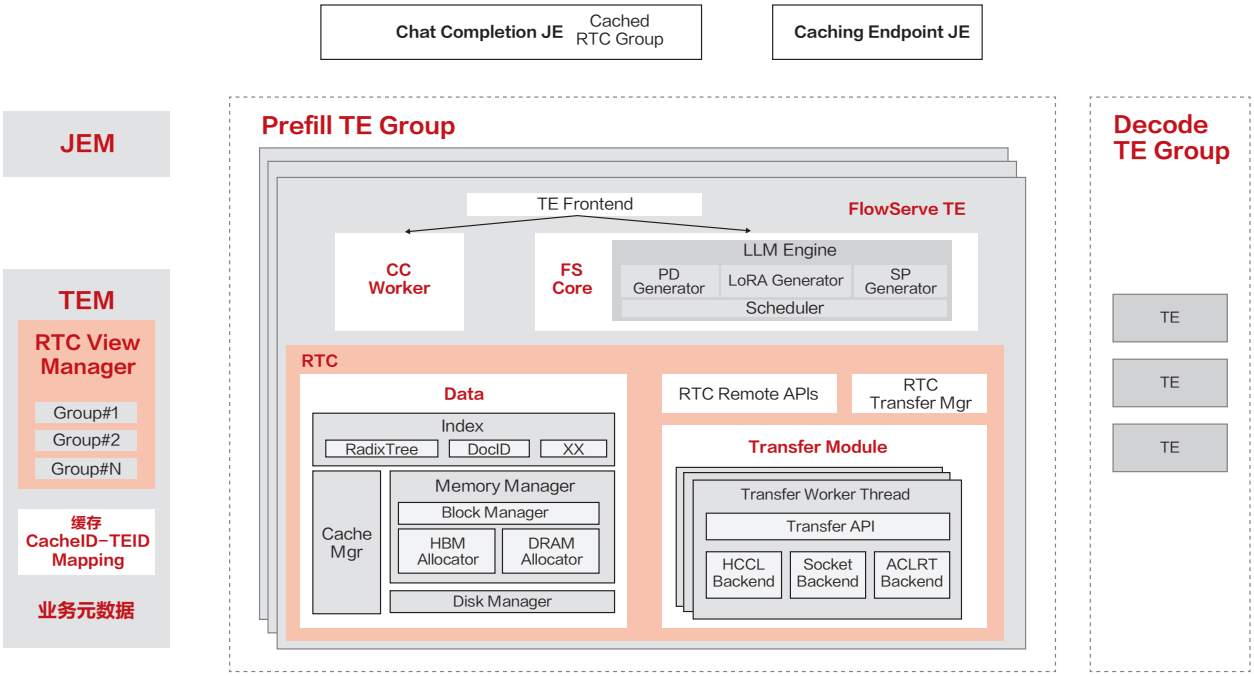


图79 PD分离组合架构

3) RTC软件架构

RTC实现3个核心功能：HBM和DRAM内存分配管理，历史KV Cache管理，以及FlowServe引擎之间的数据通信。RTC是FlowServe推理引擎的数据面和通信面；在内存管理方面，RTC有分别面向HBM和DRAM的Cache Engine；在历史KV Cache管理方面，RTC有一层混合索引支持例如Radix Tree，Session ID，Cache-ID等索引方式，以及一系列标准的前缀匹配API；在东西向数据传输方面，RTC提供了一个简单且标准的transfer Tensor API，提供FlowServe节点之间的数据传递能力。

4) 分离式推理加速：AI平台的另一个核心技术点是通过分离优化来提升服务的整体性价比。

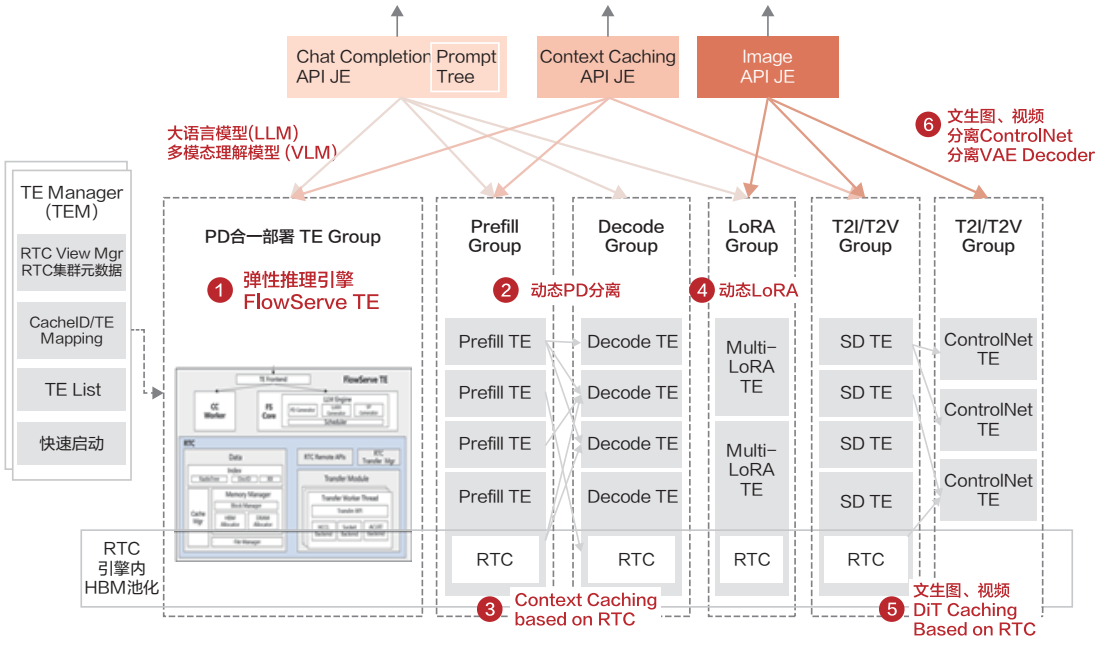


图80 MaaS推理服务的RTC架构

- 在LLM场景下，基于RTC实现的PD分离：AI平台通过RTC提供的transfer tensor API实现了经典PD分离，增量PD分离，以及D到P的数据回传等功能。如下图所示，从Prefill-only节点到Decode-Only节点的KV Cache传输通过调用RTC的transfer实现. 从Decode-only节点到Prefill-only节点的数据回传；

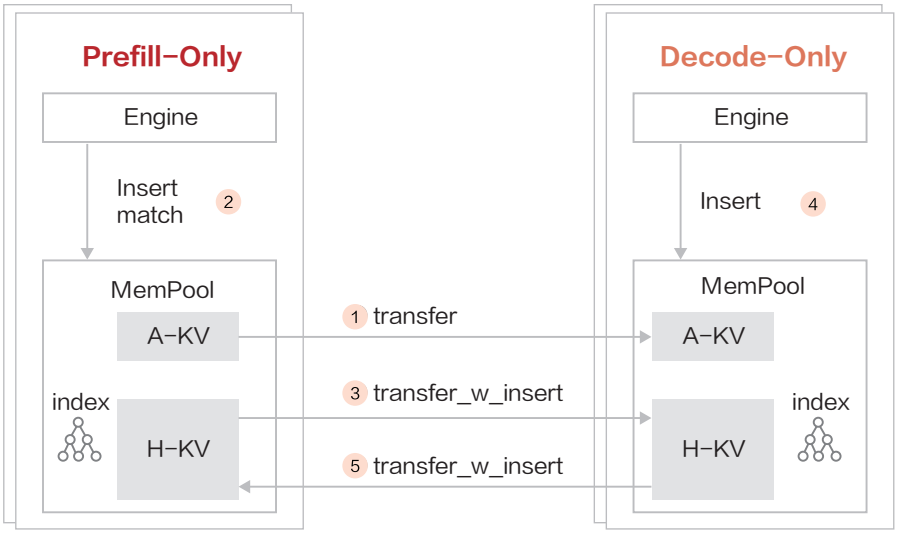


图81 LLM场景下，基于RTC实现的PD分离

- 在文生图文生视频场景下，基于RTC实现的VAE/DiT分离：在扩散模型（diffusion models）的架构中，其核心组件包括文本编码器（text encoder）、去噪模型（DiT/UNet）以及变分自编码器解码器（VAE decoder）。这些组件在资源消耗和计算需求上呈现出显著差异，这为优化部署策略提供了切入点。具体而言，text encoder虽然静态内存占用较大，但其计算量和激活量相对较小；DiT/UNet在去噪阶段则处于计算密集型状态，对计算资源需求高；而VAE decoder虽然模型静态大小较小，但激活量巨大，对动态内存需求高。鉴于这些组件在资源需求上的异质性，探索如何分离部署模型组件，并动态调整各个阶段的batching策略，对于实现资源的高效利用至关重要。特别是在文本生成图像（文生图）和文本生成视频（文生视频）等应用场景中，这种策略能够更好地匹配不同阶段对资源的需求，从而提升整体的处理效率。以CogVideoX为例，实现对VAE的分离部署，不仅能够针对性地优化资源分配，减少不必要的内存占用，还能通过动态调整batching策略，有效提升模型的吞吐量，同时降低处理延迟。这种策略的实施，对于提高大规模生成任务的处理速度和响应时间具有显著的优化效果是实现资源精细化管理和高效利用的关键策略之一。

5) 分布式调度算法：

AI 平台在调度推理请求时面临诸多挑战和高度的不确定性。这主要体现在以下两方面：（1）推理任务特性与传统负载的差异：推理请求由于底层自回归生成的特性，其执行时间可能极系统内部状态与硬件异构性，推理系统内部充满复杂的状态和多样化的硬件架构，资源分配和调度需要兼顾异构性和实时性。短也可能极长，对计算资源的需求具有高度不确定性且差异巨大。（2）系统内部状态与硬件异构性：推理系统内部充满复杂的状态和多样化的硬件架构，资源分配和调度需要兼顾异构性和实时性。要高效解决这些问题，平台需要采用负载感知和部署感知的调度算法。以下是AI平台内部现的调度算法及其功能描述：

- Cache感知算法的背景源于生成式AI模型推理过程中KV-Cache的广泛使用。KV-Cache主要用于存储模型生成序列时的中间状态，这些缓存数据在后续序列生成中会被反复访问。高效的缓存调度可以显著降低重复计算和内存读取的开销，从而提升推理性能。该算法通过在调度系统中构建分布式Token-Tree数据结构，感知KV-Cache的状态和分布，匹配缓存节点与请求之间的亲和性。通过优先将相似请求调度到共享缓存的节点，最大化缓存复用。预期可以将缓存命中率提升20%至50%，显著减少缓存失效带来的性能开销，特别适用于多实例运行和大规模推理任务。

- LoRA 感知算法针对LoRA（Low-Rank Adaptation）微调方法中的特殊问题设计。LoRA通过引入低秩矩阵避免直接调整模型权重，从而显著降低微调成本。然而，LoRA的Rank参数在硬件和任务需求上存在异构性，这种差异容易导致性能不一致或冷启动延迟。该算法感知LoRA-Rank配置，结合硬件性能特点（如 GPU 或 NPU 的适配能力），动态调整任务分配策略，避免因资源分配不均而导致的性能瓶颈。同时，通过冷启动优化机制，加速Rank初始化和预热过程。实验显示，该算法能够将冷启动延迟和首字生成时延降低23%至40%，显著提升异构硬件的微调适配效率。

- Agent 感知算法的设计旨在解决生成式AI系统从基础推理向Assistant模式的高阶推理扩展过程中面临的调度难题。高阶推理通常涉及多轮对话或任务规划，这些复杂任务的请求之间往往没有明确关联性，给系统调度增加了难度。通过感知Agent模式的上下文状态，算法对无关联请求进行逻辑分组和优先级排序，优化调度路径。同时，智能代理模块能够动态跟踪推理进度，调整资源分配策略，确保复杂任务的高效执行。实际测试表明，该算法可以将高阶推理性能提升2至3倍。

- 生成长度预测（Length Prediction）算法专注于解决生成式AI模型输出长度不确定性带来的调度问题。生成任务的输出长度具有高度不确定性，特别是在自由文本生成或长序列生成任务中，这种不确定性可能导致资源分配失衡，长任务阻塞短任务，最终影响整体系统的吞吐量。该算法通过分析历史推理数据，结合轻量化模型对生成长度进行预测，提前估算每个请求的资源需求。随后，利用动态调度机制，将长序列生成请求分散到不同资源节点，减少长任务之间的资源竞争和碎片化问题。该方法有效提高了资源利用率，优化了系统吞吐量，并显著降低了长任务对其他请求的干扰，为生成式 AI 推理提供了更加稳定和高效的性能保障。

6) Serverless冷启动：Serverless冷启动是指在无服务器架构中，计算资源在没有运行状态的情况下需要快速响应请求的过程。

在LLM推理背景下，冷启动对服务质量和响应延迟提出了更高的要求。生成式AI模型推理的计算复杂度高、资源需求大，冷启动优化直接影响系统能否高效处理动态请求和高并发任务。以下是Serverless冷启动在LLM推理场景中的关键技术点和优化策略：

- 资源快速拉起：LLM推理对硬件资源的要求极为苛刻，包括高性能的GPU、NPU或其他加速器。在Serverless架构中，为了应对突发的大规模推理请求，平台需要以最快速度启动这些计算资源。针对LLM推理场景，平台通常采用预分配计算节点池的方式，结合任务负载历史数据进行预测，提前激活可能需要的节点。此外，考虑到LLM模型推理对内存带宽的高需求，资源预热过程会提前加载常用库和模型，确保资源在拉起时达到最佳性能。
- 参数快速装载：在LLM推理过程中，模型的参数规模通常以数十亿甚至上千亿计，这些权重的装载成为冷启动的主要瓶颈。为了加速参数加载，平台采用了增量加载策略，即优先加载必要的权重满足当前任务，其他部分根据需要逐步加载。结合LLM推理特点，还会将参数存储在高性能分布式存储中，并通过RDMA等高速传输协议将权重分发到多个计算节点。进一步优化则包括使用权重压缩技术（如分块量化）和流水线解压缩，使参数加载与解压缩过程并行化，显著降低了总装载时间。这种设计尤其适合长序列推理或上下文长度较大的任务，确保冷启动延迟最小化。
- 请求快速迁移：在Serverless场景下，冷启动通常意味着将用户的推理请求从未激活的计算节点迁移到已经就绪的资源节点。对于LLM推理，用户请求的上下文和序列状态需要高效迁移以确保推理过程的一致性。平台通过引入分布式请求路由机制，能够动态检测未完成请求的状态并快速分配到合适的节点。同时，为了避免迁移过程中因序列数据丢失导致重新推理，平台会对序列状态进行实时备份并在迁移完成后快速恢复。此外，结合LLM的推理特性，迁移策略会优先考虑计算任务的上下文长度和生成阶段，从而减少长序列任务对其他任务的干扰，进一步提升迁移效率。

3.3.4 华为云AI模型OS实践

» 3.3.4.1 数智融合Serverless数据湖DataArts

DataArts Fabric是华为云推出的AI-Ready数智融合平台，旨在通过提供统一的Serverless底座，构建高效、灵活的数智融合计算与存储服务。该平台的核心是其创新的Serverless服务框架，能够自动调度和管理资源，降低开发和运维的复杂性，同时支持数据分析、大数据处理，以及AI数据工程等多种应用场景。DataArts Fabric平台通过优化计算与存储的协同，确保企业和开发者能够以最小的成本，灵活应对业务需求的变化，从而加速数字化转型进程，实现创新和竞争优势的提升。

DataArts Fabric数智融合平台产品参考如下:

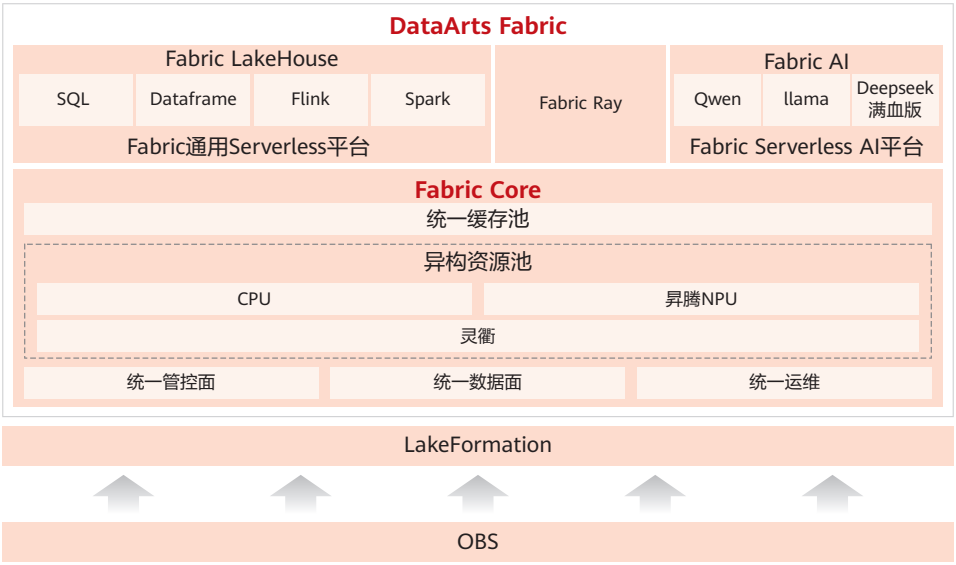


图82 华为云DataArts Fabric数智融合平台

DataArts Fabric产品基于统一的Serverless资源架构，充分发挥了Data和AI的共池调度能力，实现了CPU和NPU的异构调度。平台支持划分逻辑资源组，多产品共享同一底座，能够同时承载Data和AI负载，确保资源的高效利用与灵活扩展。DataArts Fabric平台的产品包括以下三大方面:

- Lakeformation元数据管理服务: 提供全面、集中式的元数据管理服务，旨在帮助用户高效的组织和利用存储在华为云中的各类数据资产，包括结构化、半结构化、非结构化数据，函数资产，以及模型资产等，以及对资产的管理优化和权限控制。
- Serverless平台服务: Fabric Core作为serverless平台服务，提供统一的Serverless服务框架，以及解决方案所必要的资源方案、网络方案、安全方案。并且提供高效数据方案所必须的数据缓存和数据高效流转能力，支持跨引擎复用缓存，为Data和AI负载的加速，提升整体性能。
- Serverless数智融合计算服务: DataArts Fabric的计算服务提供了多种按需付费的解决方案，能够满足不同业务场景下的高效计算需求。其核心计算能力包括Fabric Lake（包含SQL/Dataframe/AI Function/Flink Job/Spark Job等），Fabric Ray、Fabric AI等产品，支持AI和大数据分析的高效处理。此外，Fabric的各计算服务可以灵活组

合，形成高效的数据处理流水线，并支持实现了跨Data与AI作业的无缝数据流转，极大提高了数据处理效率。这些服务的组合，使得用户能够灵活地应对复杂的计算任务，提升整个数据处理和AI应用的性能与灵活性。

Fabric提供的核心功能和产品服务具体如下:

表2 华为云DataArts Fabric数智融合平台核心功能和产品服务介绍

特性名称	特性说明	应用场景
Lakeformation	支持集中式元数据管理，支持各类数据资产，包括结构化、半结构化、非结构化数据，函数资产，以及模型资产等；支持细粒度访问控制，确保资产的安全性和隐私性；支持对于数据资产的优化管理，如数据整理等服务	用于管理各类数据、函数、模型资产和对资产的权限控制
Fabric Core	提供 Serverless 服务框架，提供解决方案必要的资源池管理和弹性伸缩、网络方案、安全方案，以及数据方案（支持数据的南北缓存，和高速东西向数据流转交换）	作为产品服务的公共底座，提供服务框架、解决方案和数据方案
Fabric Lakehouse	提供 Pay-by-Use 全托管的无服务器数据湖仓引擎服务，主要用于处理和分析大规模数据集，支持结构化与非结构化数据的处理与分析、存储和管理、导入导出等；包含 SQL/Dataframe/AI Function/Flink Job/Spark Job 等功能，支持灵活的多语言自定义 UDF，并能与 Ray 等组件无缝集成	适用于传统湖仓数据的批量加工、交互式分析、报表加工等，如金融行业定时生成风控报表、电商企业动态分析用户行为日志；以及非结构化数据治理，以及 AI 大模型训练数据准备中多模态非结构化数据的清洗、加工、标注等数据工程场景
Fabric Ray	提供全托管的 Ray 服务，针对 CPU+NPU 异构算力集群，适配昇腾硬件体系，支持高性能分布式 Python 作业，降低 AI 与大数据协同开发门槛	适用 AI 大模型训练数据准备中多模态非结构化数据的清洗、加工、标注等数据工程场景，以及机器学习、深度学习、强化学习、大模型的训练和推理场景
Fabric AI	提供按需弹性的大模型推理服务，兼容业界标准 API，支持 DeepSeek、Qwen、Llama 等常用模型，支持高性价比的大模型推理、微调、模型评估	用户无需开发、部署复杂的 AI 推理和微调组件，只需关注应用开发，联网即用的 AI 大模型推理 API、微调、模型评估等场景

基于DataArts Fabric的强大能力，华为云同时推出了Serverless化的数据工程产品。该产品致力于为大模型提供从原始数据到高价值数据集的全链路解决方案，为企业高质量利用数据赋能大模型训练与优化奠定了坚实基础。该产品核心架构和功能如下图所示:

图83 华为云DataArts 数据工程产品架构示意图

产品核心的平台引擎充分利用DataArts Fabric基座，实现了统一的工作流编排与自动化执行、统一的任务高效调度、统一的元数据集中洞察与治理，以及统一的Serverless化存储与智能高效的数据流转。

在智能化数据处理能力方面，产品集成了先进AI4Data技术以赋能智能化数据处理。内置了覆盖文本、图像、音视频等全模态的60余种智能清洗算子，极大提升了数据处理效率。同时，通过包含多维度指标与检测点的标准化质量评估体系，确保了进入模型训练数据的优质性。

在此之上，产品打造了一站式数据工程平台，覆盖从数据获取、智能加工与清洗、专业标注、统一发布、精细化管理到全方位安全的一站式数据工程服务。通过可视化的数据资产地图与标准化的分级分类体系，企业能高效管理数据，并能基于模型应用效果反馈，智能优化数据配比，持续提升模型性能。

聚焦于大模型的特定训练与优化需求，数据工程产品强化了高价值数据集的生产与管理。这包括支持构建海量的预训练数据集以奠定模型广博的知识基础；高质量的微调数据集以助模型快速适配特定任务、提升专业表现；以及专业的RLHF数据集，通过细致的人工反馈提升模型的对话质量、指令遵循、有用性与无害性。这些能力共同致力于全面加速大模型训练进程并显著优化其最终效果。

这些数据工程能力，依托DataArts Fabric平台的Serverless架构共同构成了驱动大模型持续迭代优化的数据飞轮核心引擎。

» 3.3.4.2 AI开发平台 ModelArts

ModelArts是华为云提供的一站式AI开发平台，提供海量数据预处理、半自动化标注、大规模分布式训练、自动化模型生成以及端-边-云模型按需部署能力，帮助用户快速创建和部署模型，管理全周期AI工作流。

“一站式”是指AI开发的各个环节，包括数据处理、算法开发、模型训练、模型部署都可以在ModelArts上完成。从技术上看，ModelArts底层支持各种异构计算资源，开发者可以根据需要灵活选择使用，而不需要关心底层的技术。同时，ModelArts支持Tensorflow、PyTorch、MindSpore等主流开源的AI开发框架，也支持开发者使用自研的算法框架，匹配您的使用习惯。

表3 ModelArts产品形态

产品形态	产品定位	使用场景
ModelArts Standard	面向 AI 开发者的一站式开发平台，提供了简洁易用的管理控制台，包含自动学习、数据管理、开发环境、模型训练、模型管理、部署上线等端到端的 AI 开发工具链，实现 AI 全流程生命周期管理。	面向有 AI 开发平台诉求的用户。
ModelArts MaaS	提供端到端的大模型生产工具链和昇腾算力资源，并预置了当前主流的第三方开源大模型。支持大模型数据生产、微调、提示词工程、应用编排等功能。	用户无需自建平台，可以基于 MaaS 平台开箱即用，对预置大模型进行二次开发，用于商用生产。
ModelArts Lite-Server	面向云主机资源型用户，基于裸金属服务器进行封装，可以通过弹性公网 IP 直接访问操作服务器。	适用于已经自建 AI 开发平台，仅有算力需求的用户，提供高性价比的 AI 算力，并预装主流 AI 开发套件以及自研的加速插件。
ModelArts Lite-Cluster	面向 k8s 资源型用户，提供 k8s 原生接口，用户可以直接操作资源池中的节点和 k8s 集群。	适用于已经自建 AI 开发平台，仅有算力需求的用户。要求用户具备 k8s 基础知识和技能。
ModelArts Edge	为客户提供了统一边缘部署和管理能力，支持统一纳管异构边缘设备，提供模型部署、Al 应用与节点管理、资源池与负载均衡、应用商用保障等能力，帮助客户快速构建高性价比的边云协同 AI 解决方案。	适用于边缘部署场景。
AI Gallery	AI Gallery 百模千态社区，为用户提供优质的昇腾云 AI 模型开发体验和丰富的社区资源。	适用于 AI 开发探索。

ModelArts产品架构如下图:

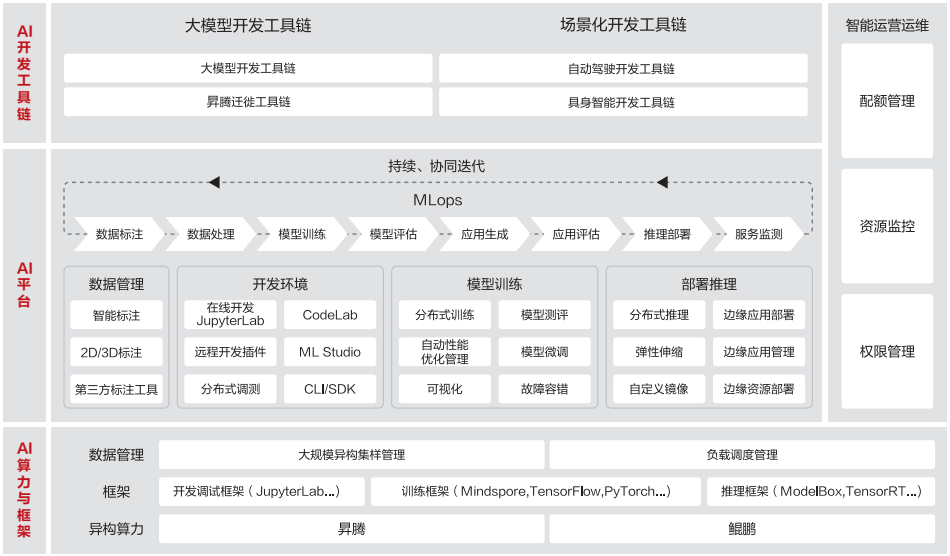


图84 ModelArts产品架构图

算力层提供全系列昇腾硬件，万卡级大规模集群管理能力，提供资源负载调度管理能力，兼容业界主流AI开发调试、训练推理框架。

AI平台层提供端到端的AI开发工具链，支持开发者一站式完成模型开发和上线，并提供高效的资源管理能力，支持自动化故障恢复，提升AI模型开发、训练、上线的全流程效率。

AI开发工具链层提供端到端的大模型开发工具链，支持主流优质开源大模型“开箱即用”，提供大模型开发套件，提升大模型开发效率并缩短开发周期。

» 3.3.4.3 昇腾云Serverless AI平台

本章节介绍华为云Serverless AI平台的整体架构，包括其API、执行模型以及关键内部模块，涉及以下几个核心维度:

- Serverless化的AI平台： 平台采用无服务器架构，用户在使用过程中无需感知底层的服务器资源配置。用户请求直接发送至平台，由平台自动执行任务调度与资源分配。Serverless的执行流程由用户请求开始，经过平台转化为内部的Job，再将Job映射为具体的Task，最后Task被分配物理资源上的Actor实例以完成处理。
- 统一的任务执行架构: Job Executor（JE）和 Task Executor（TE）。为了实现高效、灵活的任务执行，平台引入了统一的任务执行架构抽象，主要包括Job Executor（JE）和Task Executor（TE）。JE和TE分别负责Job和Task 的执行。例如，当一个推理请求到达平台时，它会首先转化为一个推理Job由JE处理。JE根据任务要求将Job拆解为一个或多个Task，然后分配到相应的TE实例进行执行。JE和TE的独立伸缩能力使平台能够灵活应对不确定的负载需求。
- 平台的控制面层。在任务执行层之外，平台设置了控制面层，以管理和协调 JE 和 TE 实例的运行状态、调度策略和扩缩容操作。控制面模块包括JE Master（JEM）、TE Master（TEM以及物理资源管理模块（PRM）。JEM负责JE实例的监控、管理和伸缩操作，TEM则负责TE实例的管理与调控；PRM负责底层物理资源的高效调度和分配，确保平台资源的合理使用。

- Model Serving模块。Model Serving层为用户提供各类模型推理服务的API接口，如Chat Completion API等，用于满足各类推理任务的需求。内部架构上，Model Serving模块由调度层、计算层和数据层组成。调度层由JE实现，负责请求的分发；计算层和数据层基于TE实现，负责具体的推理计算。在 Model Serving中，存在多种推理实例以满足不同任务类型的需求，例如基于Prefill的实例、基于Decode的实例、推理解码一体化（PD合一）实例，以及图像和视频处理实例。该模块的多层结构保证了推理服务的高并发和低延迟。
- Agent Serving模块。Agent Serving模块主要处理智能代理（Agent）相关任务的执行逻辑。该模块能够支持复杂任务的动态调用和模块组合，适合跨任务、多步骤的业务场景。在实际应用中，Agent Serving 模块可以智能地安排不同模型的调用顺序，以满足灵活、多样化的任务需求。
- 微调模块。微调模块为用户提供模型定制化的微调服务。用户可以上传小规模的数据集，通过平台进行模型的快速微调，以生成适用于特定任务需求的模型版本。平台会自动分配资源来执行微调过程，确保在Serverless架构下高效且经济地完成模型优化。

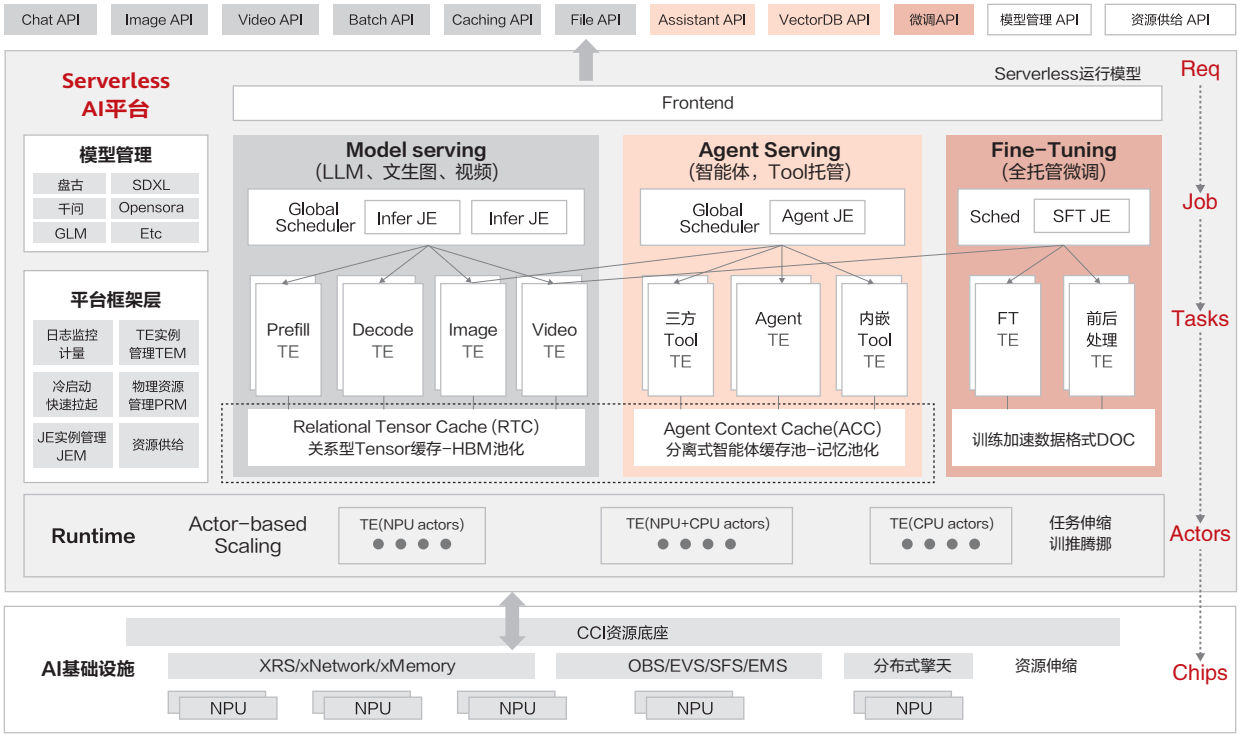


图85 Serverless AI平台架构

平台的计算范式采用自顶向下的Serverless化设计。每个请求（Request）进入平台后，会生成一个对应的 Job。在处理过程中，一个Job可能会被拆分为1至N个Task。最终，一个Task的执行可能会由1至N个Actor完成。对于平台而言，最小的伸缩单元是Actor，每个Actor对应一个底层的芯片资源，例如CPU或NPU。Serverless执行模型的设计初衷在于细化生成式AI的执行过程，实现更精细的资源伸缩能力。

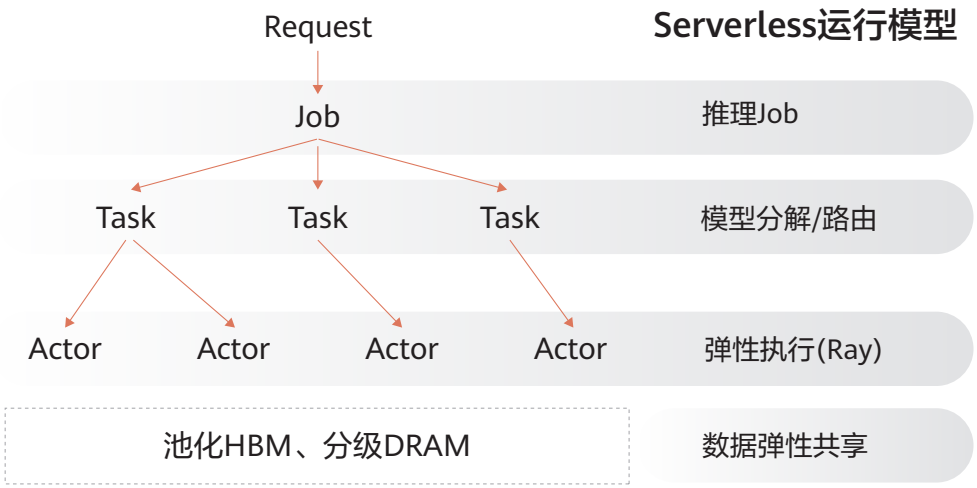


图86 Serverless AI平台运行模式

Serverless AI平台提出了一套通用的执行框架来运行所有AI负责。这套抽象围绕Request、Job、Task、Actor这几个核心概念（见上）。平台提出了如下重要抽象：

表4 Serverleress AI平台关键角色

Concept	Description
Job Executor（JE）	JE 负责处理 Job。一个 JE 是一个执行实体，处理一个类别的 Job。JE 处理 Job 时，可以把一个 Job 拆封 1 到 N 个 Task，再调用 TE 执行。JE 的主要职责是调度、编排等。举例来说，推理 JE 只处理推理 Job，微调 JE 只处理微调 Job
Task Executor（TE）	TE 负责处理 Task。一个 TE 是一个执行实体，处理一个类别的 Task。TE 处理 TaJob Exectuor Manager（JEM）sk 时，不会再拆分为更多 Task（JE 的责任）
Job Exectuor Manager（JEM）	JEM 负责管理 JE 实例的伸缩销毁等
Task Executor Manager（TEM）	TEM 负责管理 JE 实例的伸缩销毁等

3.4 AI-Naitive软硬协同优化极致性价比

在构建AI Native系统/应用时，实现算力与模型性能的极致协同需要软硬协同的深度耦合，即需要AI-Native软硬协同优化。2024年12月DeepSeek V3版本发布以来，7天用户破亿，增速超越ChatGPT，成为全球增速最快AI应用。DeepSeek大模型通过稀疏MoE、潜在多头注意力机制、多Token预测及混合精度计算等多项软硬协同策略，充分利用算力基础设施的特性，达到极致性价比（训练成本不到LLaMa 3的十分之一）。虽然，DeepSeek相关优化是面向英伟达H800的措施，但相关优化技术同样适用于昇腾NPU及其他异构AI算力平台，以及其他大模型（如盘古等）的训推加速和性价比的提升，相关技术点如下图所示。这种从算力架构到模型结构的全栈优化，最终形成软硬协同增强效应，为AI-Native系统构建出性能更高的技术基座。

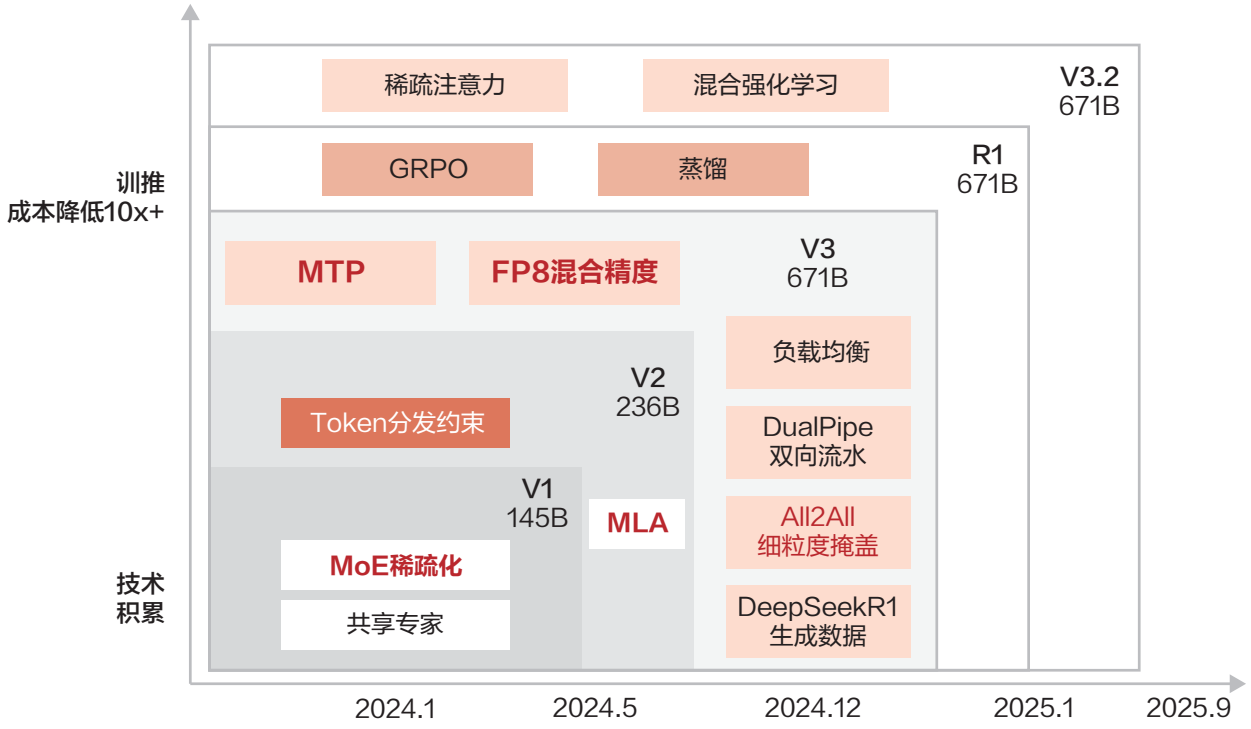


图87 软硬协同技术栈

3.4.1 大模型的稀疏MoE架构

随着模型变得越来越大，保持效率和成本成为最大挑战，尤其是在大语言模型（LLM）领域中参数扩展到数百亿、数千亿甚至数万亿的局面下。传统模型对每个输入都激活神经网络中的所有层和神经元，计算成本巨大，减慢了推理速度，并消耗大量内存。在追求效率和可扩展性的实际应用中部署如此庞大的模型是一项艰巨的任务。对此，混合专家（Mixture of Experts，MoE）被提出，MoE是一种大模型稀疏化神经网络架构。相比于类似LLaMa 3的稠密架构，稀疏MoE架构通过动态选择专门的子模型或“专家”来处理输入的不同部分，以提高模型的效率和可扩展性。这个概念可以类比为劳动分工，每个专家专注于某个大问题中的特定一小部分任务，从而更快生成结果。

MoE架构模型由三个关键组件组成：

- 专家（Experts）：专门针对特定任务的子模型。
- 门控网络（Gating Network）：一个选择器，它将输入数据路由到相关的专家。
- 稀疏激活（Sparse Activation）：只有少数专家针对每个输入被激活的方法，优化了计算效率。

门控网络充当一个选择器，它决定将哪些输入数据发送给哪些专家。不是所有专家都同时工作，而是门控网络将数据路由到最相关的专家那里。类似于 token 选择路由策略，路由算法为每个 token 选择最佳的一个或两个专家。如在下图中，输入token 1，“We”，被发送到第二个专家，而输入token 2，“Like”，被发送到第一个网络。

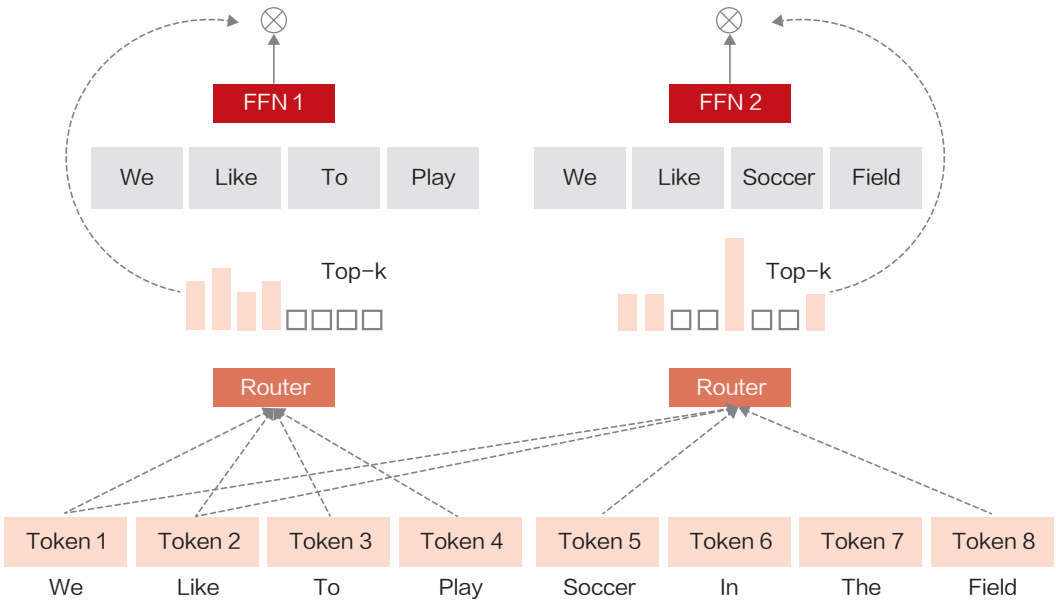


图88 MoE架构

稀疏激活是 MoE 模型的关键部分和优势之一。与所有专家或参数对输入都活跃的密集模型不同，稀疏激活确保只有一小部分专家根据输入数据被激活。这种方法在保持性能的同时减少了计算需求，因为任何时候只有最相关的专家是被激活的，大幅减少计算量，从而降低成本。不同MoE架构模型专家数量和激活率差别较大，当前国内外主流MoE模型的专家数量和激活率如下表所示：

表5 主要MoE架构模型专家数量和激活比率

模型	专家数量	激活率
GPT4	16	12% (220B/1.8 万亿)
Mixtral 8x22B	8	28% (39B/141B)
腾讯混元 (large)	16	13% (52B/389B)
Databricks DBRX	16	27% (36B/132B)
Skywork-MoE	16	15% (22B/146B)
浪潮 Yuan2.0-M32	32	9% (3.7B/40B)
DeepSeek V3	256	5.5% (37B/671B)
MiniMax-Text-01	32	10% (45.9B/456B)
盘古 Ultra MoE	256	6%
Kimi K2	384	5%

可以看出，MoE架构激活率通常5%-30%之间，每个Token只激活小部分参数，可实现极致压缩成本，相信未来越来越多的厂商会跟进MoE。从软硬协调优化的角度，单个专家所需计算和存储资源有限，需考虑把多个专家放入同一块GPU或NPU中，来节省成本。此外，对于一些专家，可能路由过来的token数量非常多，导致这些专家的计算量变大称为处理的瓶颈，这些专家可被称为热点专家。因此，昇腾云平台在大模型实际应用中，对于热点专家，可进行备份，进行冗余存储到不同的GPU或NPU中，提升处理效率，进而提升整体的性价比。

3.4.2 多头潜在注意力（MLA）

众所周知，当前大模型都是Transformer网络架构，多头注意力（Multi-Head Attention，MHA）是Transformer一个核心机制。多头注意力机制通过并行计算多个注意力头来捕捉输入序列中的不同特征。每个注意力头都有自己的查询（Query，Q）、键（Key，K）和值（Value，V）矩阵。 在实际计算中，K和V矩阵的计算开销都很大。因此，在大模型推理中，采用KV Cache缓存K和V矩阵，加速计算。然而，KV Cache是GPU和NPU显存资源巨大开销的“元凶”之一，很多情况，KV Cache占比显存用量的30%以上。当前很多对于KV Cache的优化工作。如Grouped-query Attention(GQA)，将Query分组，组内共享Key和Value参数，降低KV Cache存储空间。

多头潜在注意力MLA（Multi-head Latent Attention）更加极端，它的核心思想是通过低秩联合压缩技术，将所有的Key和Value压缩到一个向量表示，如下图所示，MLA可以大幅减少K和V矩阵的存储开销。MLA效果也十分明显，可以压缩90%的显存空间。AI-Native昇腾云采用MLA技术，可在大模型推理任务中节省1倍多的成本。

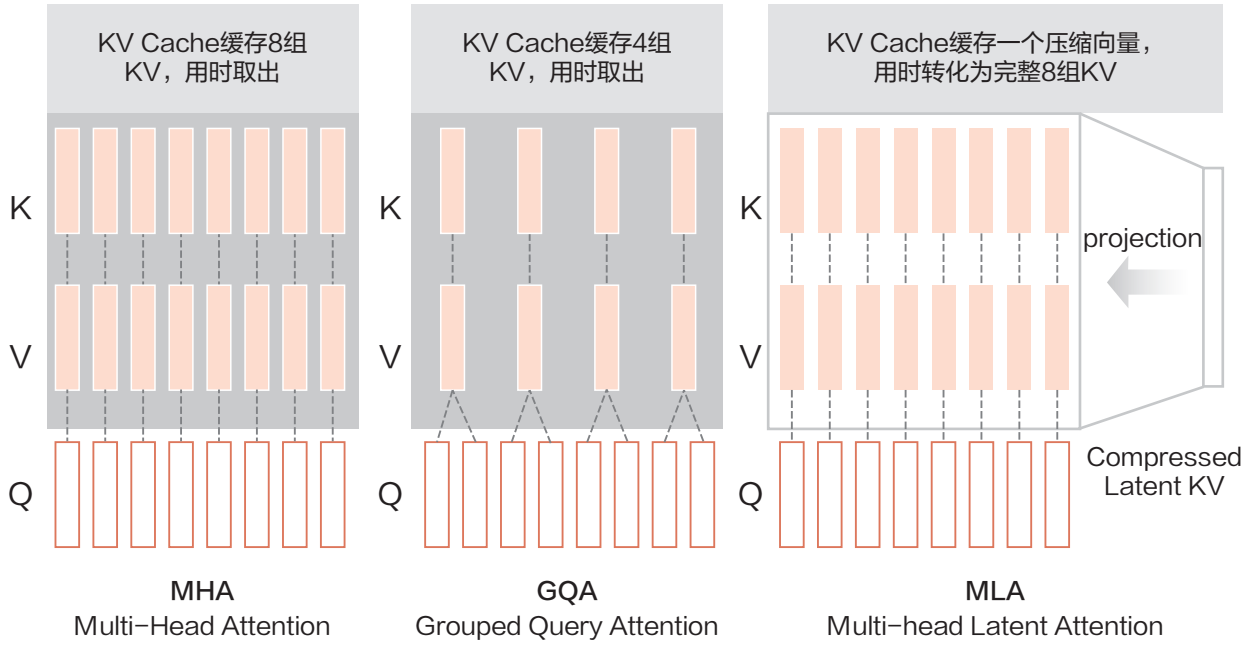


图89 KV Cache 压缩技术

3.4.3 多Token预测（MTP）

大模型预训练基本范式“预测下一个token”，Meta提出“基于多token预测”（MTP, Multi-Token Prediction）训练范式，训练时模型在每个位置上同时预测接下来的多个token，提升训练效率。如下图所示，多Token预测机制结构顺序连接了2个MTP模块，进而能够同时预测2个token，训练效率提升1倍。

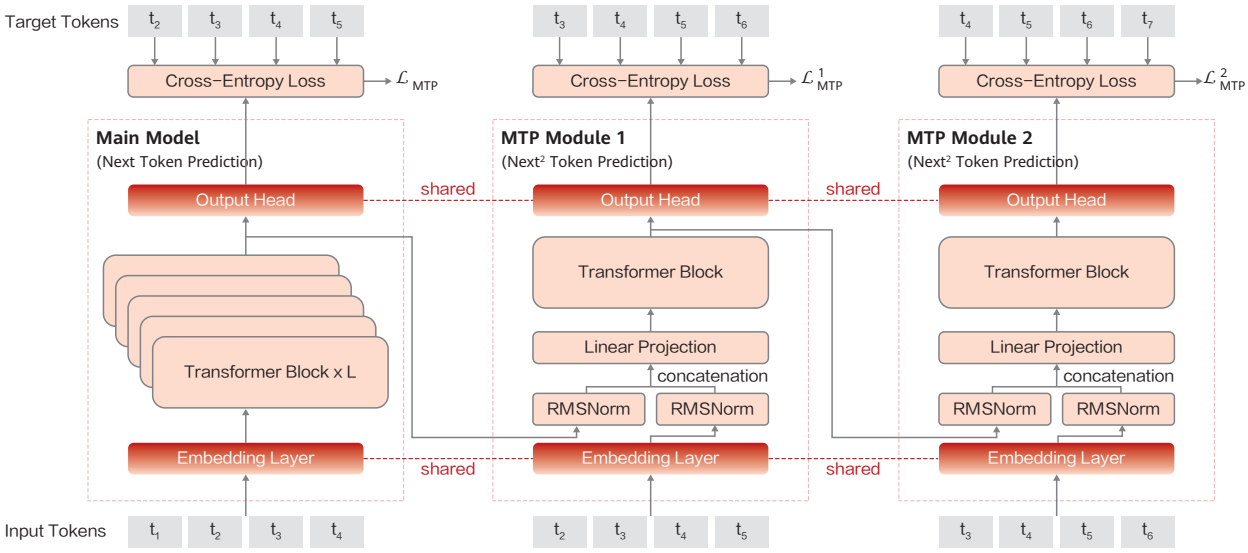


图90 MTP技术

推理阶段的MTP类似投机推理思想，是首先并行推理n个token，再对额外预测的n-1个token进行并行化验证，决定是否接受预测结果。昇腾云通过MTP机制，在预测阶段通过MTP多预测

3.4.4 极致通信隐藏

» 3.4.4.1 流水线并行通信隐藏策略

由于大模型大，单卡或单设备放不下，需切分，流水线（Pipeline）并行是将大模型中的算子切分成多个阶段（Stage），再把阶段映射到不同的设备上，使得不同设备去计算大模型的不同部分。流水线并行适用于模型是线性的图结构。如图所示，将4层MatMul的网络切分成4个阶段，分布到4台设备上。正向计算时，每台机器在算完本台机器上的MatMul之后将结果通过通信算子发送（Send）给下一台机器，同时，下一台机器通过通信算子接收（Receive）上一台机器的MatMul结果，同时开始计算本台机器上的MatMul；反向计算时，最后一台机器的梯度算完之后，将结果发送给上一台机器，同时，上一台机器接收最后一台机器的梯度结果，并开始计算本台机器的反向。

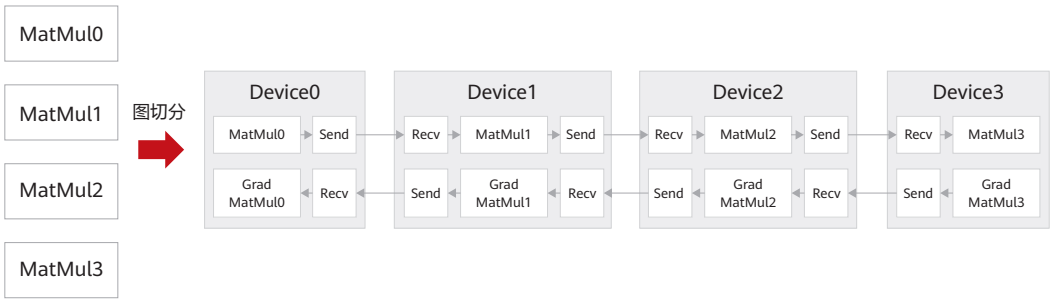


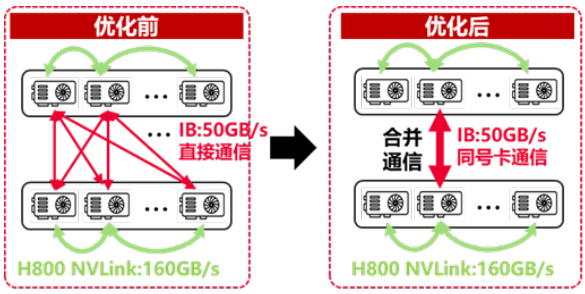
图91 流水线并行

然而，在正向计算和反向计算之间会存在大量的空闲，这种空闲称之为“气泡(Bubble)”。为了减少气泡，华为云在训练过程中，采用DualPipe流水线(双向流水线)并行的方法，大幅提升了资源利用率，进而提升性能和性价比。当前已通过优化通信算子，预期昇腾云大模型训练场景提升性能30%以上。

» 3.4.4.2 专家并行通信隐藏策略

通常情况下，在进行Mixture of Experts（MoE）的训练时，采用一种叫做专家并行（Expert Parallel）的方法。专家并行的思路是将不同的专家分配到不同的计算设备上，这有助于减少内存消耗并提高训练效率。在这个过程中，每个设备会根据MoE模型的路由规则，将自身的数据发送到相应专家所在的设备上，然后等待专家完成计算后再将结果返回到原设备。这个过程中，每两个专家之间需要进行通信，这个通信过程被称为AlltoAll通信。然而，AlltoAll通信涉及任意两个专家，通信需要的时间较长，已成为专家并行的瓶颈之一。

为了掩盖AlltoAll通信提升性能，开源库DeepEP针对H800 GPU的通信合并的优化策略，可减少20%通信开销。优化前，所有GPU直接都要建立通信，容易拥塞；优化后，节点内多个专家的运算结果合并成一份流量，再通过同号卡对外的IB发回。随着MoE架构和小专家策略的流行，专家之间通信次数变多，而每次通信量小，昇腾云可通过合并减少了节点外的通信次数，避免拥塞降低通信开销，可带来更大性能性价比收益。



3.4.5 动态负载均衡策略

MoE架构模型虽能大幅提升性价比，但是经常面临负载不均衡的问题，例如多个token指派到同一个专家处理，而这种问题在训练当中则会逐渐恶化，导致端到端性能会下降。如下图所示，MoE模型训练阶段专家的实际负载差异较大，导致性能下降10%以上。

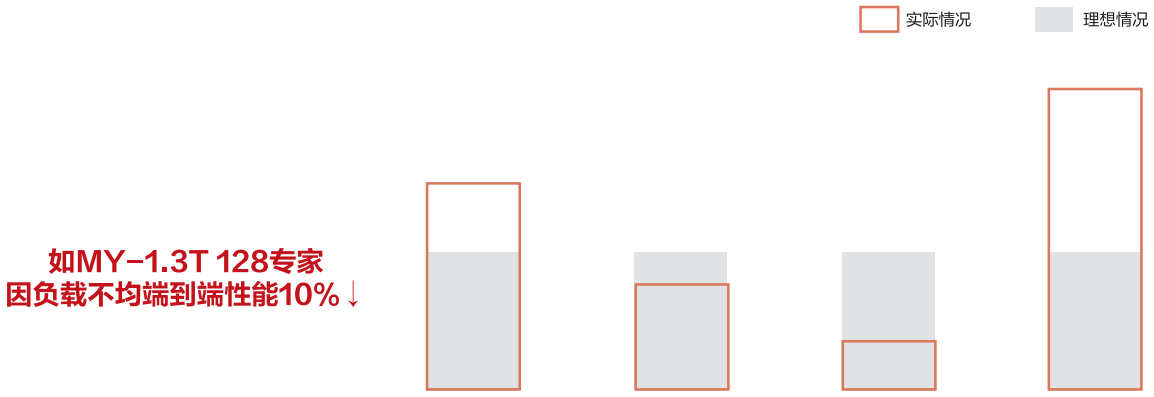


图92 多专家负载不均

为了解决多个专家之间负载不均的问题，拟采用动态负载均衡策略，通过设置偏差系数bi，调整每个专家动态负载，序列级负载均衡避免极端不平衡。如下公式所示，根据专家负载调节bi值，即若专家超载，则减小bi，若专家低载，则增加bi。

$$g_{i,t} = \begin{cases} S_{i,t}, & S_{i,t} + bi \in \text{Topk}\left(\{s_{j,t} + bi | 1 \leqslant j \leqslant N_r\}, K_r\right), \\ 0, & \text{otherwise} \end{cases}$$

此动态负载均衡策略可用于昇腾算力平台上MoE模型的训练，可能大幅提升MoE训练性能。

3.4.6 FP8混合精度计算

通常情况下，大模型的参数都以浮点数的形式存在。一般原始参数是32位浮点型，即FP32，而计算机中一个字节是8位，那么一个参数也就是4个字节大小，12B参数的模型就要占据47GB的空间。这么大的模型文件对显存要求过高，也就阻碍了大批AI技术爱好者，因此出现了量化的概念。FP16即使用16位浮点数表示，占用资源也是FP32的一半，可以大大节省资源。而BF16同样也是使用16位浮点数表示，但它8位用于指数，7位用于尾数，比FP16表示的数值范围更广。FP8以此类推，使用8位浮点数表示参数。

相比于FP16或BF16，采用FP8低精度浮点数训练，显存、带宽等资源占用仅为原来的一半，可大幅提升训练的性价比。为了使训练精度不降低，可在精度敏感算子（embedding、gating、attention等）保持16位精度，而在计算密集型的算子(前向计算、反向激活/梯度矩阵计算)均采用FP8计算，既保证了精度，又可提升性能。具体混合精度训练流程如下图所示。 通过实际实验测试，使用FP8混合精度训练与BF16训练误差维持在0.25%以下。

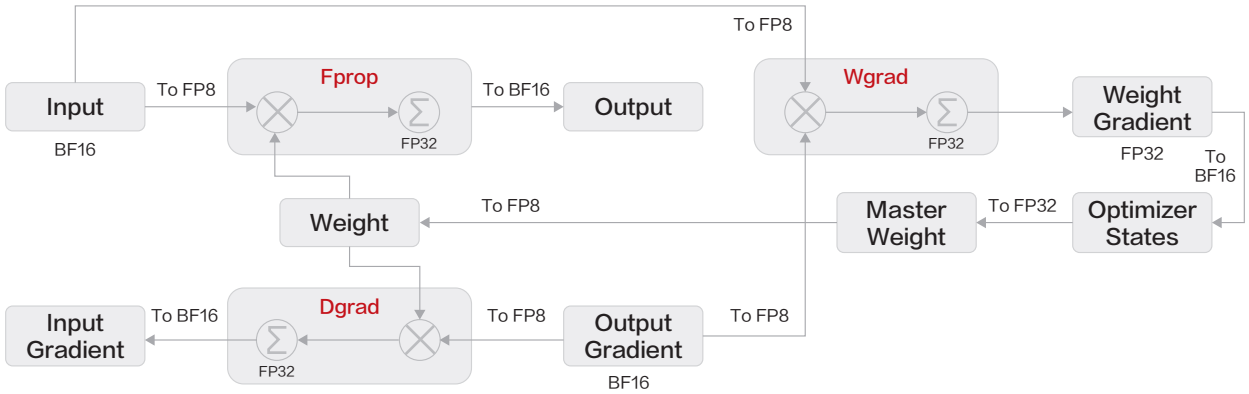


图93 混合精度计算

3.4.7 昇腾云软硬协同优化实践

华为昇腾云CloudMatrix384超节点量身定制了一套软硬协同的大模型推理服务系统CloudMatrix-Infer。它为部署大规模 MoE 模型（如DeepSeek、盘古 Ultra MoE等）提供了最佳实践。CloudMatrix-Infer 最核心的创新之一是其点对点服务架构，它将 LLM 推理工作流解耦为三个独立的功能子系统：预填充（Prefill）、解码（Decode）和缓存（Caching），简称PDC解耦，如下图所示。

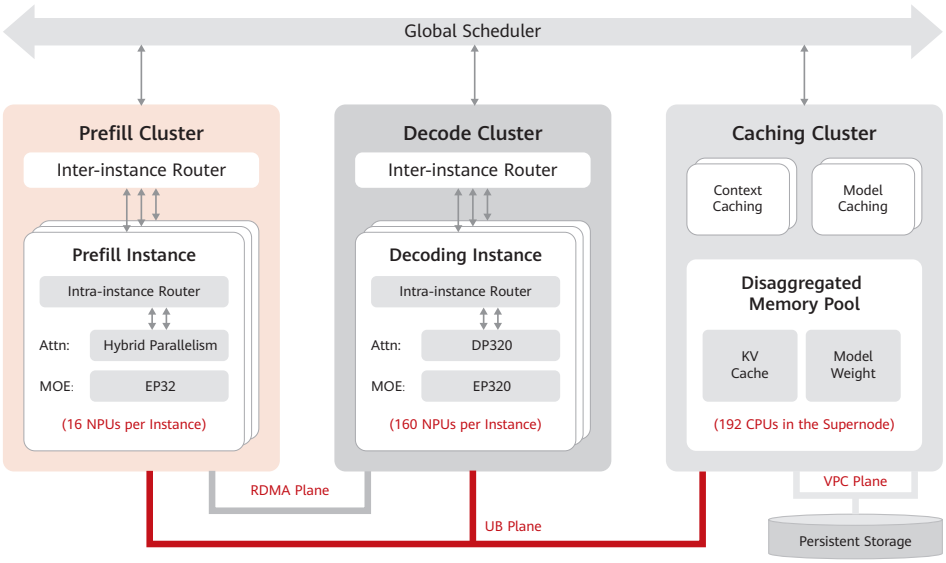


图94 CloudMatrix-Infer PDC架构

相比于现有推理系统，CloudMatrix-Infer采用PDC架构的优势如下：

- 利用UB网络优势。CloudMatrix384 的超高带宽 UB 网络使得远程访问不再是瓶颈。所有 NPU（无论是预填充还是解码 NPU）都能直接、以统一的带宽和延迟访问一个分布式缓存集群（基于解耦内存池）。
- 解耦调度与数据局部性。请求调度不再受限于KV缓存的物理位置。推理请求可以被分发到任何可用的NPU实例，大大简化了调度逻辑，显著改善了系统范围内的负载均衡和 NPU利用率。
- 统一弹性缓存。通过汇聚预填充和解码节点上的DRAM资源，系统形成一个统一、弹性的缓存层，提高了内存利用率和缓存命中率，更好地应对突发或倾斜的工作负载。
- 独立可扩展。预填充和解码阶段的性能瓶颈不同（预填充通常计算密集，解码通常内存带宽密集）。解耦允许根据实际工作负载需求，独立地扩展预填充、解码和缓存集群的资源，实现更精细的硬件分配。

基于DeepSeek在英伟达H800硬件平台上的软硬件协同优化经验，CloudMatrix-Infer针对昇腾NPU的架构特性，将在下面章节详细阐述其定制化软硬协同优化技术的具体实践方法。

1) MLA 优化实践

预填充阶段的 MLA 计算是主要瓶颈。DeepSeek 在 H800 上采用纯数据并行（DP），但在昇腾 NPU 上效率不高，原因在于：

- 序列长度偏差：实际请求输入序列长度不同，导致 DP 下 NPU 负载不均衡。短序列 NPU 早早完成，然后空闲等待。
- 并发不足：如果并发请求数少于 DP 度，部分 DP 分片会空闲。

而CloudMatrix-Infer的混合并行方案： 通过在不同阶段切换并行策略，解决长序列和并发不足带来的负载不均衡问题。具体来说，将MLA计算分解为三阶段，并应用不同的并行策略，如下图所示。

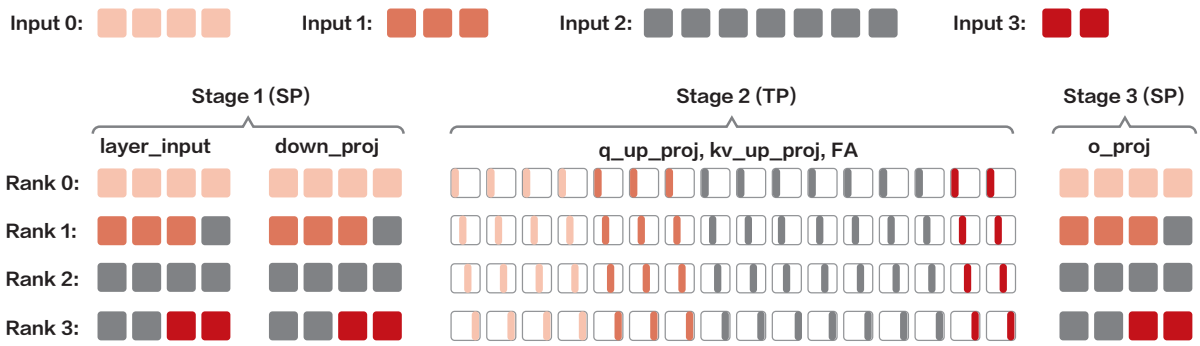


图95 MLA混合并行策略(SP-TP-SP)

第一阶段（输入处理、down_proj）和第三阶段（o_proj）：采用序列并行（Sequence Parallelism，SP） 结合序列打包。将多个请求的序列拼接在一起，然后将打包后的超长序列段分配给不同的 SP rank。这样，无论原始请求长度如何，所有 NPU die 都能获得大致均匀的 token 数量，实现负载均衡。

第二阶段（q_up_proj，kv_up_proj，FlashAttention 核心）：采用张量并行（Tensor Parallelism，TP）。由于 MHA 计算的每个注意力头是独立的，可以将注意力头均匀分布到不同的 NPU die 上，提升资源利用率。

数据重分布：在不同并行策略之间切换时，通过All-Gather（第一阶段到第二阶段）和All-to-All（第二阶段到第三阶段）操作进行数据重分布，减小通信开销。

经过MLA优化，CloudMatrix-Infer推理性能相比于英伟达H800硬件有1.2倍左右的提升。

2) MTP 优化实践

DeepSeek采用多Token预测（MTP）提升推理性能。然而，DeepSeek的MTP的实现常因频繁的CPU-NPU同步而效率低下。因此，如下图所示，CloudMatrix-Infer对MTP进行了以下优化：

- 聚合元数据初始化：在解码步骤开始时，预先计算并批量处理所有动态元数据张量（如序列长度），并直接存储在 NPU 内存中，消除 CPU 的重复干预。
- 无CPU干预的NPU内采样：将整个采样过程（token概率排序、累积和计算、候选过滤）迁移到NPU上，并融合到MTP和LLM验证图中。这消除了CPU-NPU之间的交互，使得计算图可以在NPU上背靠背连续执行。

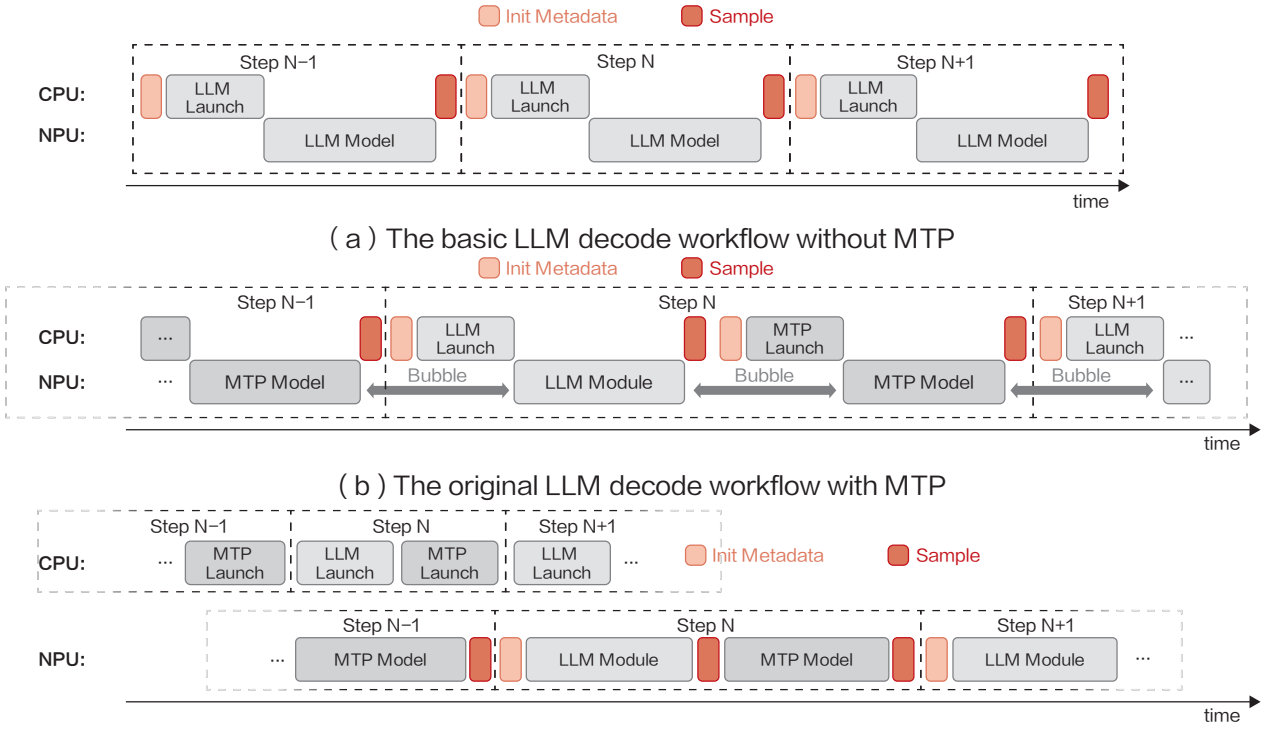


图96 MTP优化实践

这些优化消除了CPU-NPU协调导致的流水线中断，实现了NPU的持续并行执行，MTP性价比提升30%。

3) 通信优化实践

MoE模型解码的关键在于token分发（Dispatch）和专家输出组合（Combine）这两个通信密集型步骤。传统的MoE实现存在通信开销大、动态形状导致效率低、串行依赖等问题。CloudMatrix-Infer设计了FusedDispatch和FusedCombine两个融合算子来解决这些问题。CloudMatrix-Infer的融合算子优化如下：

表6 融合算子优化策略

AIV-Direct 直写	SDMA 通信启动慢，低延迟场景成瓶颈	让 AI Vector 核心直接写远程 NPU 内存，跳过 SDMA	极大降低启动延迟，通信更轻量
早期量化	通信数据大（BF16 格式），带宽压力大	发送前将数据量化为 INT8（附缩放因子），减半消息体积	通信数据量减小，带宽占用骤降
共享内存预分配	动态分配 + 同步开销大，序列数变化引发波动	预先为所有通信对手分配好缓冲区，静态图执行	消除动态分配、CPU-NPU 同步开销
数据发送流水线	目标偏移计算与数据发送串行导致等待和空闲	设计三阶段流水线，计算、量化、传输并行重叠	完美重叠计算与通信，隐藏延迟

4) 混合精度实践

在混合精度方面，将16位运算(FP16/BF16)量化为FP8对于提升性能和性价比意义重大，但现阶段昇腾平台暂不支撑FP8计算。昇腾云平台计划通过8位整型（INT 8）替代FP8进行运算，可带来的好处如下:

- 计算效率提升: 低精度数据使计算更快。
- 内存占用降低: 8位整数仅为16位浮点数内存的一半，节省HBM资源，提升性价比。
- 内存带宽需求降低: 数据传输减少，带宽瓶颈缓解。

如何在降低精度的同时，最小化对模型准确率的影响？CloudMatrix-Infer 设计并实现了一套训练无关、分层（Hierarchical）的 INT8 量化方案，在昇腾 910C 平台实现高性能推理，同时谨慎管理精度损失。

表7 混乱精度策略

策略名称	主要思想	实现方式	主要优势
混合精度策略	关键路径用 INT8，敏感部分用 BF16/FP32	FFN/Attention 用 INT8，归一化 / 门控等用 BF16/FP32	性能与精度兼得，优化整体推理效率
自适应缩放因子搜索	最优缩放因子映射浮点到 INT8，减少量化误差	离线搜索最佳缩放因子，误差最小化	离线完成，无运行时开销，误差低
异常值抑制与结构转换	抑制长尾 / 异常值对量化精度的影响	量化前线性变换（旋转 / 缩放吸收），均衡数据分布	异常值影响降低，误差不放大
高效 INT8 矩阵乘法核	INT8 量化需高效硬件内核支撑	优化 GEMM 内核，激活值 token 级量化，权重通道级量化，内存布局优化	大幅提升吞吐，充分发挥硬件计算能力
块级裁剪与误差补偿	针对局部权重特征优化精度	权重分块独立裁剪，插入误差补偿项	精度提升，无需再训练，部署快速

此外，在昇腾硬件架构创新中，华为提出了一种新型数据格式HiF8，其核心是引入动态位宽机制，通过灵活调整指数域与尾数域的长度分配，使数据表示更贴合实际计算需求。相较于传统FP8格式，HiF8显著扩展了数值表达范围，接近FP16的动态范围，有效避免了极端值计算中的溢出风险；同时相比INT8整数格式，HiF8在低精度计算场景下可保留关键数值特征，提升精度。这一创新数据格式已深度适配昇腾下一代AI芯片架构，为高能效大模型推理和训练提供了兼具宽动态范围与低比特精度的算力支撑。

3.5 AI-Native技术赋能的云服务

华为云基于“AI for Cloud”理念，将AI-Native技术赋能云服务智能升级，使得云服务更为智能、便捷、高效。

3.5.1 软件开发生产线CodeArts盘古助手

软件开发生产线CodeArts盘古助手是基于盘古研发大模型的智能开发助手，以研发大模型为核心的IDE插件，覆盖CI场景，提升增量代码开发、存量代码优化、开发者测试效率；协同CD场景，提供超级入口支持一站式应用部署，重塑了智能化软件研发的新范式，让开发者更加聚焦业务创新。

CodeArts 盘古助手关键特性如下：

- 高质量训练数据：基于业界开放数据、华为内部研发数据调优训练，模型更加精准。
- 上下文工程：基于用户输入，结合内部工程化经验，补充多维度背景，最终传给大模型的内容上下文信息越完整，可获取答案越准确。
- 多模型调度：开源、商用、自研等多模型融合调度，用户场景推理效果最大化。
- 更准确：代码生成在HE评测同等参数模型业界领先，代码有效接纳率达到20~30%。
- 更深度：深入而精准的研发问答，项目级上下文/RAG/SFT等一站式解决方案。
- 全流程：软件开发读、写、调、测、查等编码全过程智能辅助。

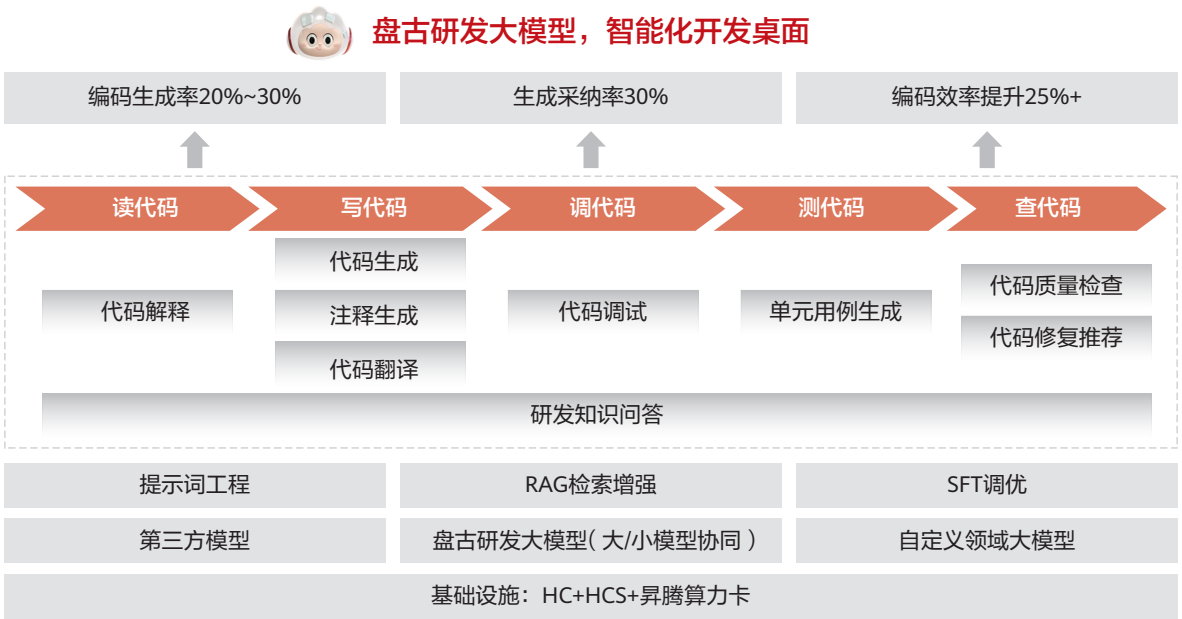


图97 盘古研发助手

CodeArts 盘古助手具备智能生成、智能问答两大核心能力，覆盖如下开发场景：

- 代码生成：根据自然语言生成完整代码逻辑，实现编码效率的大幅提升。支持Python、Java、C、C++、JS、TS、CSS、Html等语言；支持CodeArts IDE、PyCharm、IntelliJ IDEA、VSCode等IDE。
- 研发知识问答：对话框内进行任意研发相关的知识提问，提高研发问题的解决效率。
- 单元测试用例生成：自动创建单元测试用例，提高测试覆盖率，确保每个功能和场景都被测试到。
- 代码解释：快速分析代码并自动生成文档，提高阅读代码的速度和效率。
- 代码注释：快速分析代码，自动生成有意义的完整注释，增加代码可读性，提高同一个代码库注释风格的一致性，提升研发问题解决效率。
- 代码调试：根据运行代码报错提示的错误日志，一键修复代码。
- 代码翻译：快速分析代码并自动完成迁移和翻译，提高开发者工作效率。
- 代码检查：快速分析代码，提供错误详细信息和修复建议，缩短错误定位时间。

3.5.2 安全云脑盘古助手

随着云计算和数字化转型的加速，云环境日益复杂，企业安全管理面临不断演变的威胁。多云、混合云架构和微服务普及导致安全边界模糊，传统安全方式对弹性与动态云资源的适应力不足。同时，攻击手段的自动化与智能化带来更高频、更隐蔽的风险，零日漏洞和内部威胁频发，考验着安全体系的快速响应和威胁预测能力。人工分析效率低、规则库有限、响应滞后等传统痛点迫切需要新型安全能力的支撑。

AI-Native 云安全应运而生，它将人工智能技术深度融入云安全管理与防御之中，实现智能识别、自动化运营、自适应防御和预测性安全防护。这一新模式不仅降低了安全技术门槛，更为安全运营带来智能化变革。

安全云脑盘古助手，即云脑安全大模型，正是AI-Native云安全理念的代表。作为融合AI与安全的创新平台，它集成了自然语言理解、自动分析、任务编排等多项智能能力，重塑云端安全运维方式，赋能企业快速、高效、精准应对安全挑战，为云脑平台带来以下核心价值：

- 效率提升：通过自动化分析和研判，将复杂安全事件的平均处置时间缩短至3分钟以内，整体效率提升15%~20%，让安全团队从繁琐重复的工作中解放出来。
- 精准识别：基于海量数据和深度学习，准确识别各类安全威胁，例如攻击行为、恶意脚本等，有效降低误报率，确保安全防护的准确性。
- 智能决策：提供智能化的处置建议和研判报告，为安全团队提供决策支持，帮助其快速、准确地应对安全事件，减少潜在损失。
- 能力增强：降低安全分析的技术门槛，即使没有深厚安全知识背景的人员也能快速上手，提升团队整体安全防护能力。

图98 安全云脑盘古助手应用场景

云脑安全大模型由多个核心子模型和智能中枢组成，围绕智能分析、研判、识别与自动化执行展开：

1) 四大核心安全模型

- 攻击研判和处置大模型：自动识别攻击类型和风险级别，形成处置建议和研判报告，覆盖主机、应用、网络等主流安全场景，单事件处置用时少于3分钟，准确率85%以上。
- 攻击识别大模型：专注恶意脚本检测与分析，可高效识别主流恶意Bash脚本类别，大幅提升安全响应速度。
- 攻击溯源大模型：自动分析告警、事件、漏洞和资产信息，降低安全数据分析门槛，无需复杂SQL即可深入洞察344项核心数据字段。
- 安全领域大模型：精细分析主机和四/七层载荷流量，为安全团队提供全面、深入的攻击溯源和防御建议。

2) 智能中枢能力

- 意图理解模型：依托大模型自然语言理解能力，精准识别用户提出的各类安全需求，覆盖来源于知识问答、安全事件分析、攻击识别等六大主要场景，意图识别准确率高达99.9%。
- AI Agent智能体：具备根据任务自动调用安全工具、动态编排任务的能力，实现安全运维的自动化和高效协同，支持多工具两跳调用，任务执行准确率95%。

云脑安全大模型的核心在于将大模型（LLM）和AI Agent智能体深度融合，引入自然语言理解、自动化安全分析与任务智能编排等前沿技术。这些关键技术赋能云安全体系实现高度智能、自适应和主动防御，成为推动AI-Native云安全落地的核心驱动力。

1) 大模型（Large Language Models，LLMs）和安全应用场景结合:

- 意图理解：大模型通过对海量安全数据的学习和训练，能够精准理解用户的安全意图，无论是安全知识问答、安全事件的分析、还是安全工具的查询，都能够快速准确地响应。例如，用户可以用自然语言提问："分析一下最近的告警事件"，大模型就能理解并自动进行相关分析，并给出研判报告。
- 攻击溯源：自动分析云脑平台产生的告警、事件、漏洞和资产信息，降低Flink SQL的使用门槛，帮助安全人员快速定位攻击威胁，为后续的研判和处置打下基础。
- 恶意脚本攻击识别：专门用于识别各种恶意脚本，并输出恶意性分析报告，辅助安全人员进行研判，及时阻止攻击行为，提升安全防御能力。
- 攻击研判&处置：结合云脑平台中的安全数据，大模型能够自动分析告警、事件、漏洞等信息，快速研判攻击类型和风险程度，并提供处置建议和研判报告，大幅缩短安全事件的响应时间。
- 自动化安全策略生成：大模型能够根据企业自身的安全需求和云环境的特点，自动化生成安全策略，并根据实际情况进行动态调整，提高安全防护的有效性。

2) AI Agent 智能体智能调用安全任务编排

- 智能安全助手：AI Agent 能够理解安全人员的指令，自动执行安全任务，如漏洞扫描、配置检查、事件响应等，极大地提高安全运营效率。
- 动态任务编排：AI Agent 可以根据安全事件的优先级和依赖关系，动态编排安全任务，并自动调用相关安全工具，实现安全工作的自动化。例如，当检测到恶意代码攻击时，AI Agent 可以自动调用漏洞扫描工具、日志分析工具等进行联合分析和处置。
- 安全工具集成：AI Agent 可以集成多种安全工具，实现安全工具的有序调用和协同工作，将不同工具的能力进行有效整合，提供更全面的安全防护。

AI-Native云安全是云安全发展的必然趋势。它不仅能够解决传统云安全面临的挑战，还能够为企业提供更智能、更高效、更主动的安全防护。随着AI技术的不断发展和成熟，AI-Native 云安全将在未来发挥越来越重要的作用，成为企业数字化转型的重要保障。

3.5.3 数据库盘古助手

AI-Native数据库通过全面融合AI算法、工具等技术，重塑数据库业务全流程，提升用户体验。华为云数据库通过融合领域大模型、场景小模型、智能体框架等关键能力，实现了华为云GaussDB Doer数据库盘古助手，构建全流程的AI维护和数据库开发辅助能力。

华为云GaussDB Doer数据库盘古助手具备纵向和横向两个能力维度：

- 纵向系统能力方面，聚焦数据库产品自身的AI-Native能力及特性功能，包括在内核调优、系统诊断、领域问答提供AI辅助。华为云数据库在数据内核，已经实现并应用了基于规则及代价的优化器，同时实践并推进基于AI的优化器落地；以此作为基础，在面向用户的数据库领域套件层，提供具备自适应优化的SQL优化智能体套件，提高数据分析和SQL开发效率。在系统层面上，利用AI全流程诊断，构建跨组件、全链路的系统自动、自主感知处置能力，实现“1分钟感知、5分钟诊断、10分钟恢复”的确定性数据库智能运维能力；同时提供面向数据中心级的数据库智能问答题，低门槛掌握了解数据库运行运维实践、运行状态等信息。
- 横向数据库使用全流程方面，在涉及到数据处理的全流程提供智能辅助，全流程包括规划设计、开发测试、迁移优化、运维管理、数据洞察等阶段，利用AI4DB的能力，极大的提升了全流程开发和运维效率。

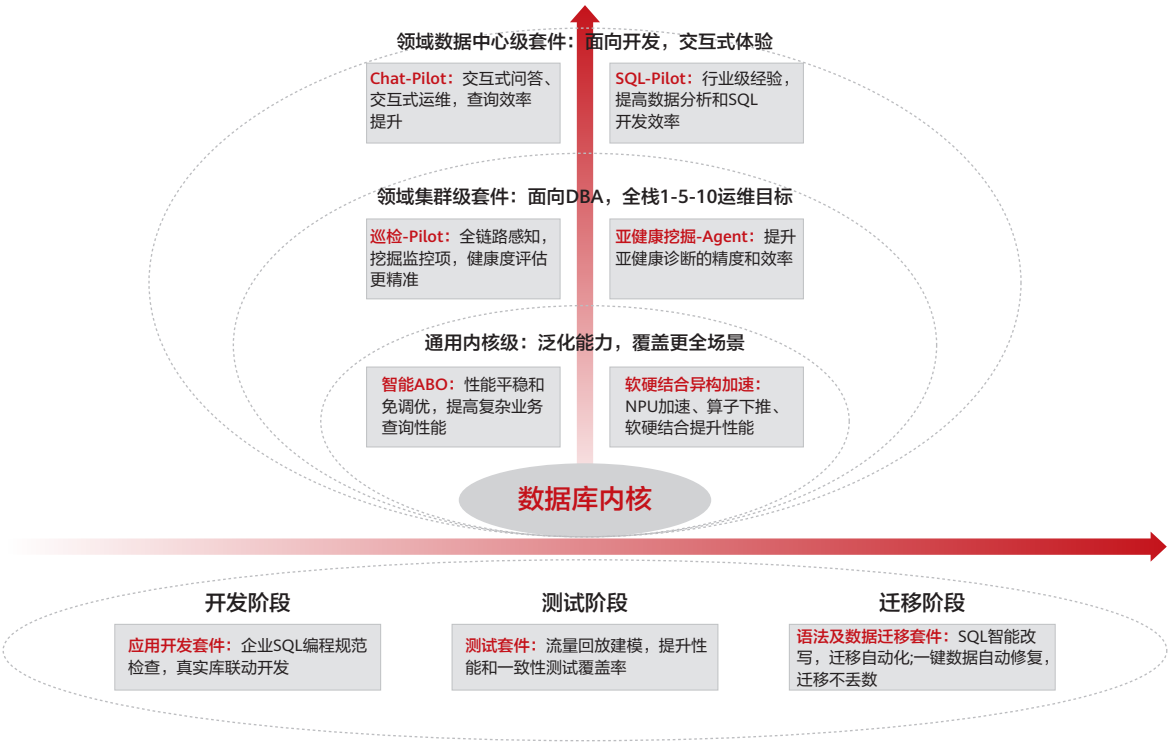


图99 数据库盘古助手架构概览

云数据库GaussDB Doer数据库盘古助手提供了“数据库问答智能体”、“SQL优化智能体”及“数据库诊断智能体”三个关键场景的服务能力，融合大模型任务规划和知识增强能力实现数据库全链路保障：

1) 数据库问答智能体

数据库问答智能体面向客户使用数据库中整个生命周期涉及到的知识、案例、实践提供交互式问答的查询服务。面向不同的数据库用户，通过自然语言输入、多轮交互的方式，快速准确地提供数据库知识结果。用户通过问答智能体反馈的知识结果，即可完成数据库的功能评估，数据库安装、配置，数据库特性功能应用实践方法，运维处理等事务。

华为云数据库通过三层技术能力支撑数据库问答智能体实现易用、准确的问答体验：

- 交互式上下文问答：在用户通过自然语言进行提问输入后，通过语言大模型进行语义解析理解，对原始查询语句进行规则改写，提升检索知识命中率；同时结合AI算法赋能的意图识别以及会话记忆的上下文召回匹配的检索内容修正，进一步提升准确率。在通过数据库问答智能体的检索查询到匹配的知识内容后，结合知识库进一步进行回答内容的生成及校验，从而使得追问场景问题交互信息更精准，结果答复场景输出回答指导用户问题可闭环。
- 融合检索查询优化：数据库问答智能体在进行知识检索查询时，通过假设性回答、向量及文本的多路融合检索及检索后重拍审核的关键能力，提升答复结果的准确率。
- 自更新的数据库知识：数据库产品功能繁多，知识的领域专业性以及时效性成为了挑战。华为云数据库问答智能体基于华为云数据库自身十余年最佳实践积累，构建了数据库垂域大模型，同时形成了包含数据库功能使用、部署、运维等各类场景上百万已提炼结构化的专业知识库，确保专业性及完善性。在知识的时效性上，问答智能体具备自动化的知识更新能力，针对不同模态、输入格式（文档、社区、图片等），通过华为云盘古大模型能力进行私有域知识的自动提取以及知识库的更新和融合。

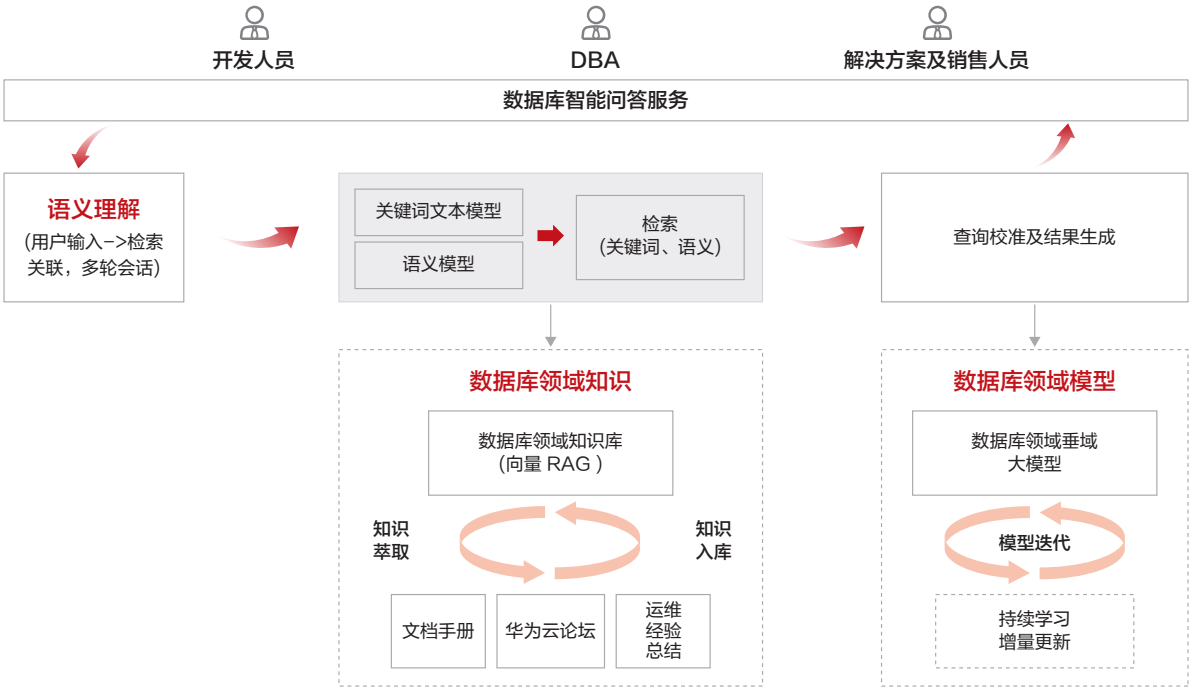


图100 数据库问答智能体架构

通过上文数据库智能体的关键技术能力，在开发、运维、解决方案技术分析上为客户带来价值：对开发人员，SQL规范、环境配置、调用返回信息解释，开发效率更高；对运维人员，知识即经验，有效支撑运维操作，运维检索更准确；对解决方案销售人员，提供数据库关键特性、架构比对收益等等信息，快速匹配需求，方案推荐更精准。

2) 数据库运维智能体

在数据库运维场景，面对故障排查流程长（5~10步）、处理效率低（小时~天级），业务深入后数据库规模增长对DBA经验技能越来越高要求的挑战，华为云数据库提供具备自主智能的运维辅助工具提升运维效率，下文将具体描述对应的关键特性能力：

- 一键式的数据库状态感知及根因分析：华为云数据库通过构建运维Agent及运维大模型控制中心，基于实时日志及关键指标监控数据，进行系统运行风险预测及智能巡检。在识别到指标偏离、跳变等非预期状态，通过自编排的根因分析及故障图谱进行故障自复盘，锁定故障根因，为DBA运维人员推送风险告警及分析报告，快速掌握运行态势。
- 主动识别及自诊断：在发生异常运行状态的场景中，往往存在大量告警等信息。通过大模型及不同特性Agent融合，结合思维链（TOT）以及RAG知识库进行异常诊断及故障恢复手段的匹配识别，进一步结合恢复风险预测、备机模拟验证等手段，在运维人员决策评估安全可控的前提下，进行故障的自动恢复，实现故障的快速恢复。

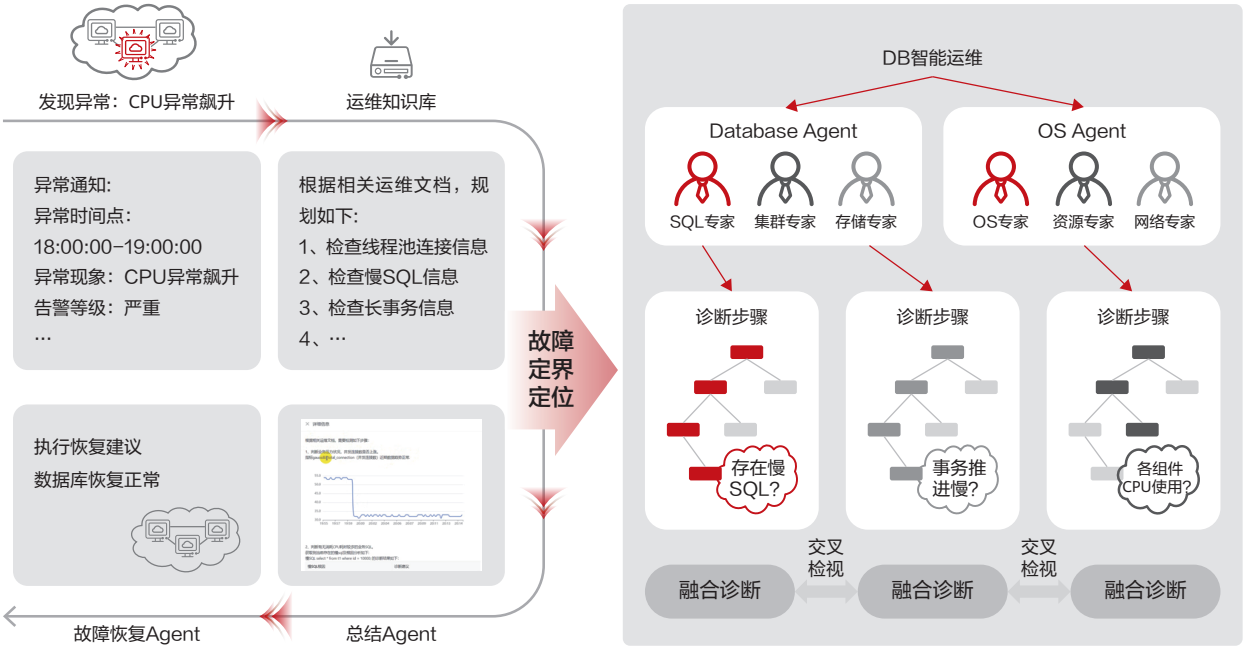


图101 数据库运维智能体架构

在华为云数据库的日常运维工作及客户化的实战场景中，数据库运维智能体支持典型的故障场景分析及自动化的自诊断能力，实现故障场景MTTR 50%的降低，保障生产安全。

3) SQL优化智能体

在数据库的使用场景中，SQL语句的输出质量以及在运行系统中的稳定运行对应用的稳定十分重要。华为云数据库SQL优化智能体，在SQL开发及运行运维的全生命周期提供全链路监控及之智能优化能力。开发阶段，通过基于SQL语法规则的静态审核及基于动态执行计划于历史执行性能SQL特性智能匹配的手段，识别烂SQL，并给出优化建议，避免引入烂SQL ；在运维运维阶段，通过识别慢SQL，并给予大模型识别慢SQL的优化方案进行持续的在线优化，确保系统运行持续稳定。其中的关键技术能力如下:

- SQL自动检测识别。对用户输入的SQL，通过生产库历史SQL模板进行聚类，结合历史执行性能提取特征，快速识别烂SQL，避免异常SQL引入。
- SQL自动优化。针对慢SQL及已识别的烂SQL，结合历史专家优化经验、SQL优化构建大模型等AI能力，实现SQL优化改写。同时通过专家经验SQL优化知识库，大模型结合知识增强能力，能够精准返回优化解决方案。

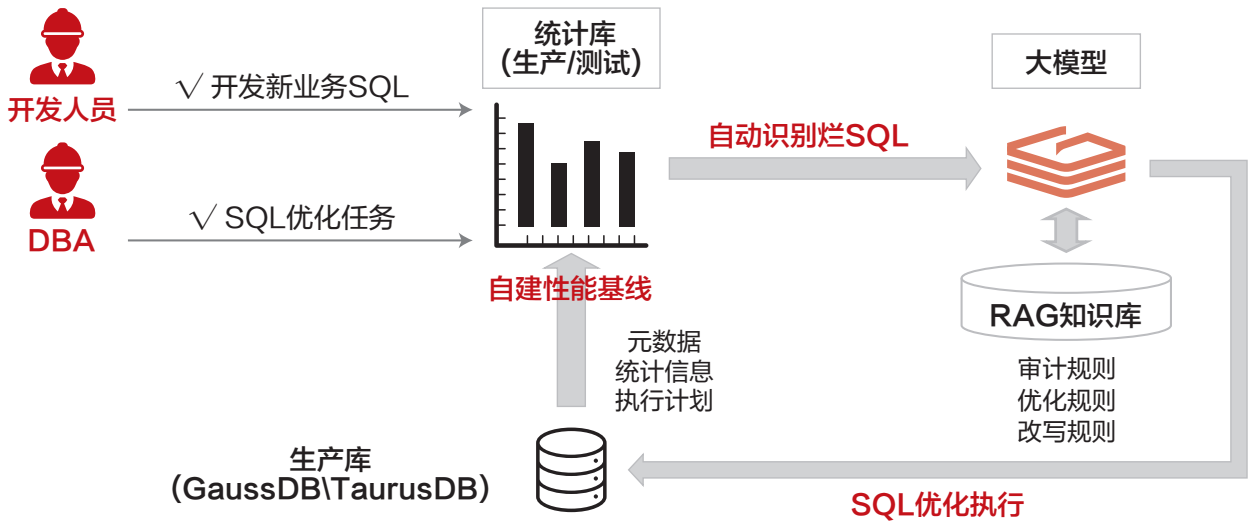


图102 数据库SQL优化智能体架构

通过数据库SQL优化智能体，能够准确识别并有效避免系统引入烂SQL，同时通过自动化、智能化的SQL优化，使数据库用户SQL执行效率倍增。

3.5.4 数据治理生产线盘古助手

在AI-Native的数据治理方面，由传统的结构化数据治理快速向融合数据治理转变，应用AI技术驱动数据治理的自动化和智能化。在数据治理场景下，AI和ML的深度应用引入了自动化和增强各种数据相关处理过程的范式转变，主要包含如下方法：

- 自动化与智能化：使用AI技术可以自动化数据处理任务，例如数据准备和数据清洗。借助ML算法可智能识别数据存在的重复和不一致等问题，并进行自动修复，减少人工干预；利用NLP技术，自动进行数据的分级分类和标签化，提高数据的可管理性和可检索性。在技术实现上，根据要处理场景的实际需要，灵活采用AI小模型或大模型算法或技术。
- 预测性分析：在AI技术的帮助下，通过分析历史数据，可以对未来趋势或潜在问题做出预测。使用AI Workflow对确定性的业务进行工作流编排来驱动业务决策；通过AI和ML模型，对数据进行深度分析，预测潜在的问题和趋势，为数据治理策略的制定提供智能决策支持。
- 增强用户体验：AI重塑数据治理E2E流程，AutoETL自动准备数据，Copilot辅助生成SQL等代码、作业和开放API，面向未来，AI算法可根据用户的需求和行为提供个性化的支持，进一步提升用户体验。
- 扩展数据治理范围：随着AI技术的发展，治理的数据范围由结构化数据扩展为包含多模态数据在内的全域数据，帮助用户实现实时数据处理和分析，企业能够实时监控数据的生成和使用情况，动态地实施数据治理策略，确保数据在快速变化的环境中仍然能做到符合质量和安全标准。
- 构筑知识语义层：在企业的数据和业务之间，构建一个知识语义层，提供知识服务，驱动业务侧分析、洞察、搜索和智能决策等业务应用，也可为大模型提供可用于微调和预训练的领域知识数据。

AI-Native数据治理通过全面拥抱AI，显著提升数据治理的自动化和智能化水平，增强了对全域异构数据的治理能力，提升了实时数据的处理与分析能力，通过知识语义层连接企业的数据和业务并为AI模型提供领域知识数据。一方面是AI使能数据治理效率提升和数据价值释放，另一方面是数据治理为AI提供更优质的数据，两者相辅相成。

在AI-Native的数据洞察方面，随着基于大模型的数据洞察技术的兴起，企业现在能够更加直观和高效地从大量数据中挖掘出有价值的洞察。这一技术的核心在于利用人工智能（AI），特别是自然语言处理（NLP）能力，实现了数据查询、分析和报表制作的自动化和智能化。下面，将探讨AI-Native数据洞察在企业中的三个主要应用场景，展示它如何帮助不同角色的员工在各自的职能领域实现数据驱动的决策。

- 领导层的决策支持：在快节奏的商业环境中，领导者需要迅速获得关键业务指标（KPIs）和运营数据，以便做出及时的策略调整。AI-Native数据洞察技术允许领导通过自然语言查询来直接获取这些数据，无需深入了解复杂的数据查询语言或依赖数据分析团队的报告。例如，一位公司高管可以简单询问：“这个季度的销售额比去年同期增长了多少？”系统能够即时解析这一查询，自动提取相关数据，并生成易于理解的报告。这不仅大大缩短了信息获取的时间，还使领导能够基于最新的数据洞察做出决策，优化公司的经营和运营策略。
- 业务人员的取数需求：对于业务人员而言，及时了解自己的业务指标完成情况及其背后的详细数据至关重要。AI-Native数据洞察能够帮助他们通过简单的自然语言指令，了解个人或团队的KPI完成情况，并进一步探索影响这些指标的因素。比如，销售人员可以询问：“本月我的销售目标完成率如何？”系统会提供一系列相关指标，包括销售额、目标完成率、与上月或去年同期的比较，甚至建议哪些行动可以帮助达成未完成的KPI。这种即时的反馈和建议，不仅提高了个人和团队的业绩，还有助于精细化管理业务流程。

- 数据分析师的高效报表制作：数据分析师在传统的工作流程中，需要花费大量时间在数据查询及报表制作上。通过自然语言处理能力，可以极大地提升这一流程的效率。数据分析师可以直接用自然语言描述他们想要的报表内容和格式，系统则能自动理解这些需求并生成相应的报告。这不仅释放了数据分析师的时间，让他们能够专注于更复杂的分析任务，还提高了整个组织的数据分析能力和效率。

AI-Native数据洞察通过其先进的自然语言处理能力，为企业各层级员工提供了一个强大的、直观的数据访问和分析工具。从领导层到一线业务人员，再到数据分析师，每个人都能够更快速、更有效地从数据中获取洞察，支持他们的决策和行动。随着这种技术的不断发展和完善，可以预见，数据驱动的决策将成为企业文化的核心，帮助企业在竞争中脱颖而出。

数据治理生产线全面落地盘古助手，下文中针对数据治理平台DataArts Studio和数据洞察平台DataArts Insight两个云服务的实践展开描述。

DataArts Studio的智能分析助手，能力贯穿数据治理端到端全流程，聚焦三大场景：数智开发场景、数据治理场景、数据分析场景。

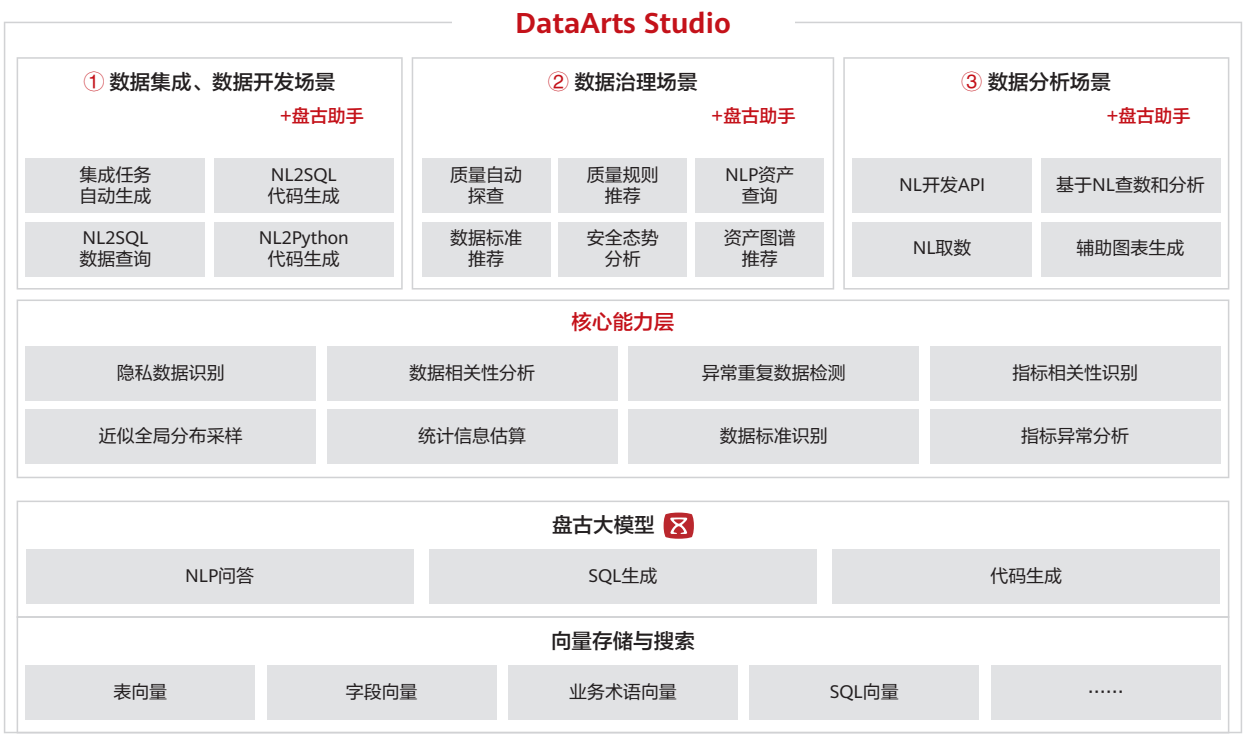


图103 DataArts Studio智能分析助手架构示意图

其中，数智开发场景，优先落地AI辅助生成SQL，包含Hive SQL、Spark SQL、DLI SQL和DWS SQL，接下来是AI辅助Python代码生成和数据集成作业生成，SQL的解释、注释和优化、测试等功能逐步构建；数据治理场景，助手辅助数据质量自动探查、数据质量规则智能推荐，数据地图NLP搜索和资产图谱推荐，数据安全安全态势智能分析等；数据分析场景，基于自然语言实现自动分析、找数和用数，并辅助生成图表(简化图表直接生成和展示，复杂图表与DataArts Insight协同)。

使用LLM重塑数据治理全流程，智能助手助力管数用数效率翻倍：

1) 数据管理智能化

- AutoETL数据集成自动化
- 结合知识生成SQL准确性提升2倍
- 自动数据质量管理
- 数据资产图谱智能推荐

2) 数据管理模型核心能力增强

- 业务词库和业务指标管理增强语义理解
- 基于基础大模型优化，泛化能力更强
- 大模型+小模型结合，模型效果更佳

3) 知识湖

DataArts Studio的知识湖，连接企业的数据和业务，构筑起企业知识语义层。整体架构如下：

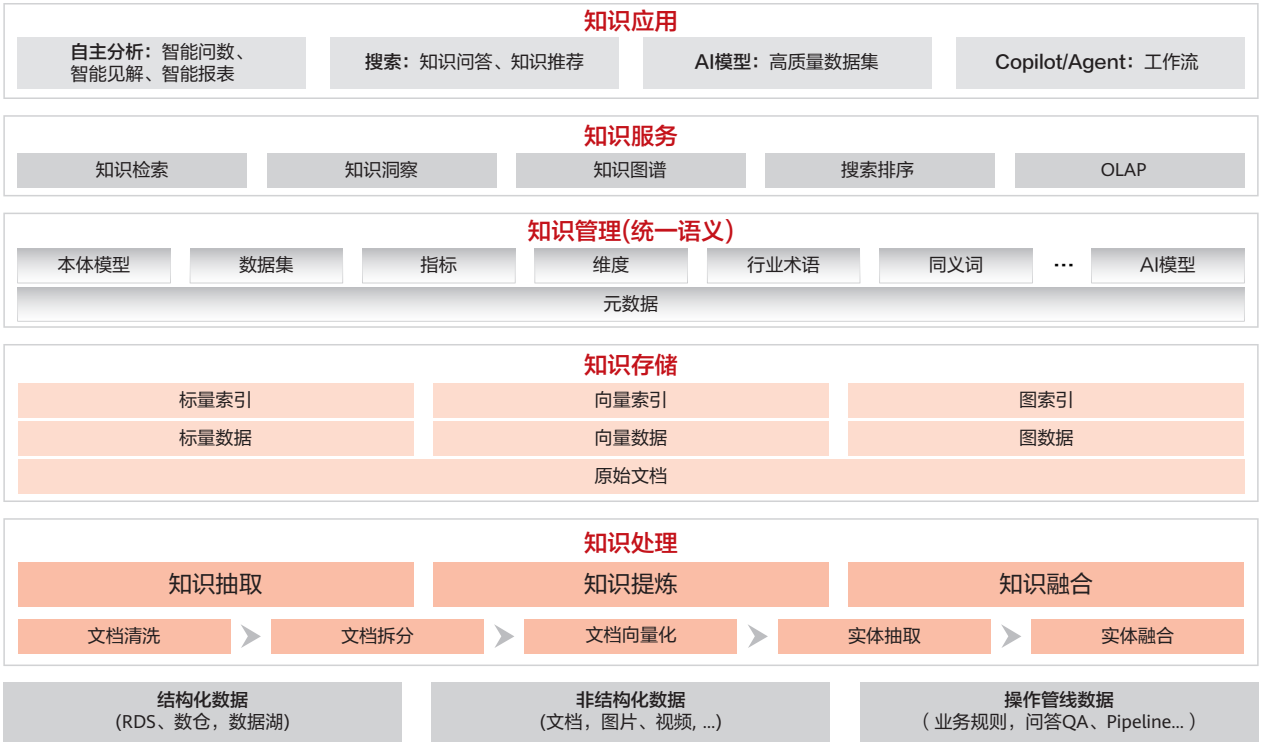


图104 DataArts Studio知识湖架构示意图

DataArts Studio为企业构筑以知识为中心的数据底座，用知识湖来升级传统的数据湖，通过数智融合应用激活企业数据价值，驱动企业业务智能决策。架构上分层如下：

- 知识处理层: 知识处理层对企业内部各类数据(结构化/非结构化数据)进行抽取、提炼和融合，最终得到知识。
- 知识存储层: 存储已经处理好的知识数据，主要包含标量、向量和图谱数据及其索引数据。
- 知识管理层: 知识管理层统一语义，配置生成语义模型、指标、维度、层次、同义词和行业术语等。自动生成元数据，实时自动捕获业务逻辑。
- 知识服务层: 对外提供知识服务，支持RAG、GraphRAG、MemRAG随机访问; 支持OLAP(SQL/Text2SQL)统计分析，并能为大模型提供领域模型Fine-tuning数据集。
- 知识应用层: 知识驱动上层业务应用，主要包括自主分析场景的智能问数、智能见解和智能报表; 搜索场景的知识问答和知识推荐; AI模型场景的高质量Fine-tuning数据集供给; Copilot和Agent场景的智能决策等等。

企业通过构筑知识湖，可实现多模态数据的自动化知识抽取、融合、索引，最终形成统一的语义层，高效连接企业数据和业务。

DataArts Insight智能分析助手的整体架构分为三层:

第一层是盘古NLP大模型（L0），该模型做到读万卷书，通过对文本与代码数据的训练，具备基本常识和推理能力。

第二层是BI大模型，BI大模型是基于NLP大模型训练出来的，专门为数据分析服务，能够更加准确高效的处理一系列数据分析任务，如: NL2SQL，NL2JSON，多轮对话，Data2Insight。并通过知识增强，数据泛化，思维链等技术，来优化模型的效果。

最上层是BI大模型应用层，智能分析助手基于BI大模型开发出一系列应用，如指标查询，智能见解，AutoGraph，自助取数，报表查询。

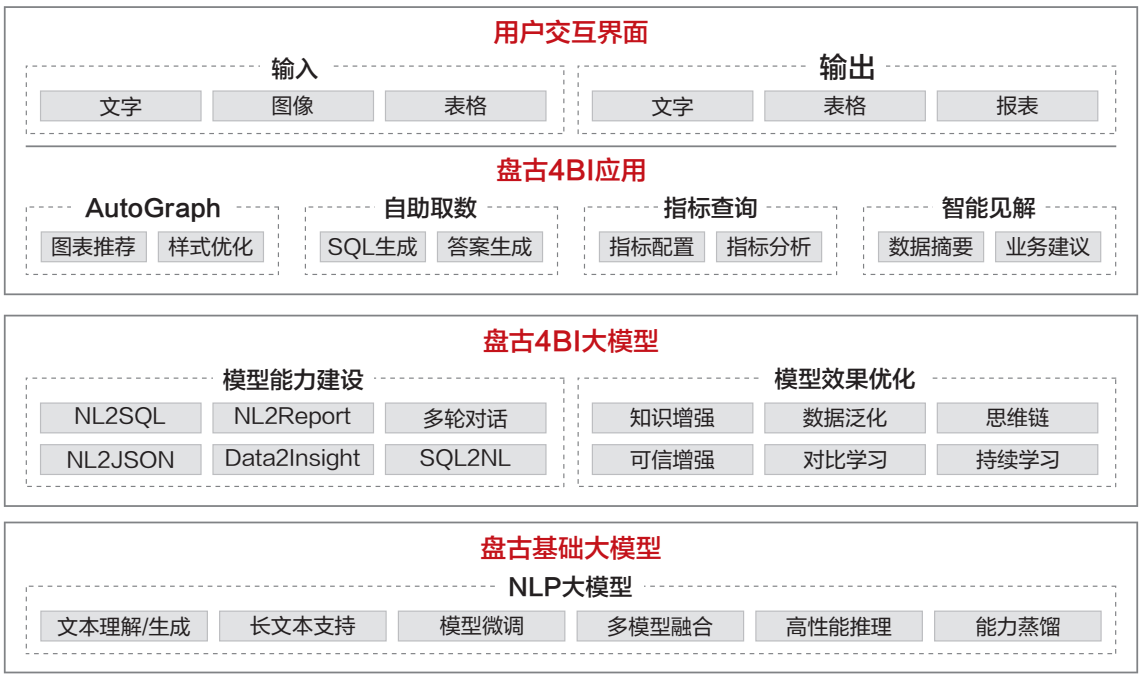


图105 DataArts Insight智能分析助手架构示意图

智能分析助手的技术方案分为两个阶段:

- 配置阶段是由技术人员来完成。类似于报表的制作过程，首先连接数据源（如: DWS，Clickhouse），然后创建数据集（可以理解为视图），数据集上面可以定义字段名，字段描述，同义词，表之间的连接关系，基于数据集可以创建指标，包括原子指标、衍生指标、复合指标，最后创建问答领域，将所需要查询的数据集放到问答领域，进行训练。训练的过程是把元数据信息，同义词，指标，业务专业词库做embedding，插入到向量库。
- 问答阶段面向的是最终问答用户，提供自然语言接口，供用户查数据和分析数据。给定用户的自然语言查询，智能分析助手首先会对用户的问题进行多轮改写，然后将问题输入到语义检索模块，该模块会讲与用户查询相关的表和列检索回来，构造成一个prompt发给BI大模型，BI大模型生成DSL，发给DaaS引擎，DaaS将DSL转成最终可执行的SQL，SQL执行完以后得到的结果会通过AutoGraph模块完成可视化，最终调用BI大模型生成见解。用户可以对结果进行进一步的分析，如: 异常检测，归因分析，趋势预测等。

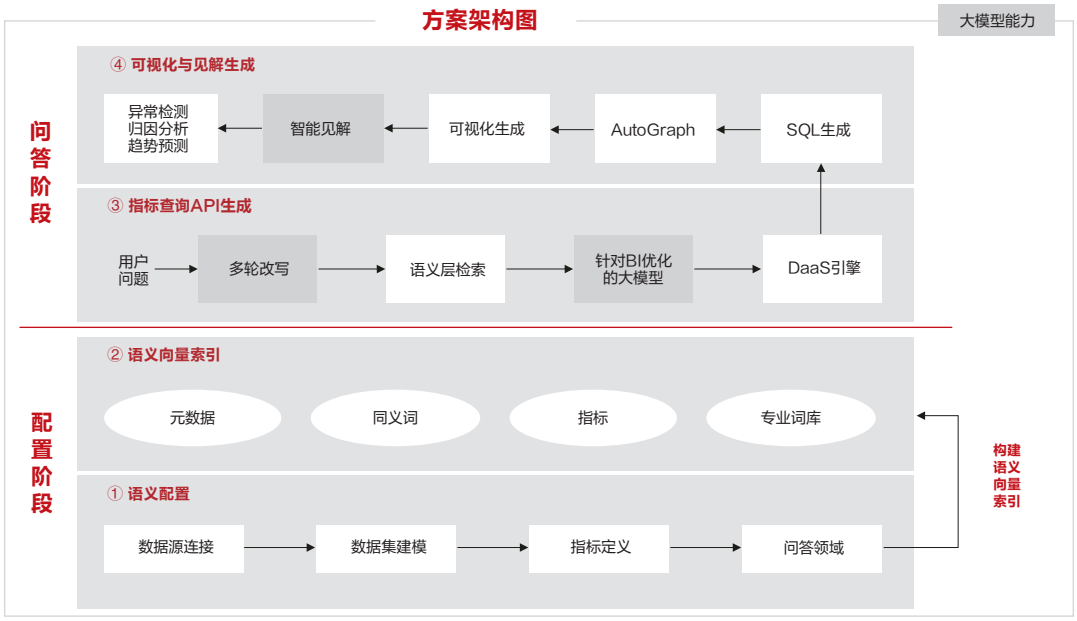


图106 DataArts Insight智能分析助手技术方案示意图

该方案具备以下竞争力:

- 全面的知识灌入: 支持数据集建模（多表，宽表），元数据注入（字段，表，枚举值），各类指标定义（原子，衍生，复合），同义词，专业词库等全面的语义配置能力，让大模型BI更懂客户的业务知识。
- 高准确率: 针对BI进行特别优化的大模型，提升自助取数的准确率。
- 结果可信: 提供查询结果解释，对于不确定的查询结果基于告警提示，并可以对结果进行修改，实现结果可信。
- 数据安全: 提供数据行列权限的配置，自动将此权限自动将此权限加到最终生成的SQL，让用户可以通过大模型安全的访问敏感数据。
- 超低门槛: 通过在提问的时候对用户问题自动补全，在返回结果后对下一轮问题进行推荐，降低用户问什么以及描述问题的门槛。

3.5.5 云运维盘古助手

华为云在运营与运维方面积累了丰富的经验，形成了以确定性运维为核心的理念，通过构建确定性恢复能力，实现故障的快速发现、定界定位与恢复。然而，随着AI Native技术的兴起，传统的运维模式已无法满足日益复杂的智能化需求。因此，华为云积极探索AI技术在运维领域的应用，以提升运维效率和质量，运维大脑因此产生，运维大脑是数字化平台运维领域深度使用AI的应用和产品，它结合了大/小模型的协同、人机协同等技术，基于service on service 的架构思想，基于华为云的能力构建，实现运维领域的AIOps智能化。

如下图，早期的运维大脑基于华为数十年的运维知识积累，结合运维领域的垂域大模型和专属小模型以及运维工程。

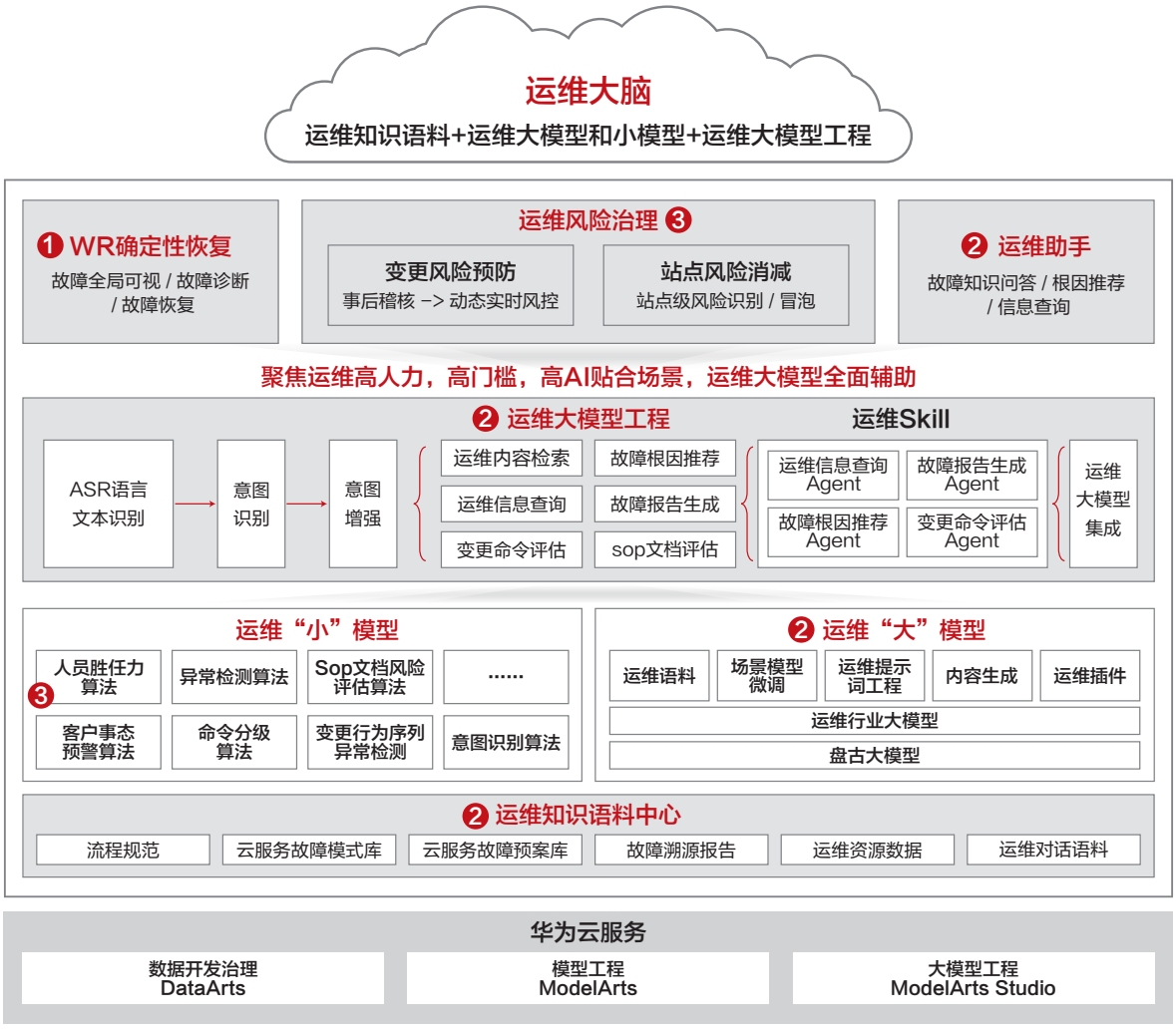


图107 早期运维大脑架构

主要覆盖以下三个能力:

- WarRoom确定性恢复：“人+平台”的方式为运维人员，通过端到端的数据联通和治理，提供故障的全局可视，故障点风险诊断和故障恢复能力。

- 运维风险治理：通过Sop文档风险评估算法，命令分级算法、变更行为序列异常检测等算法实现运维工作的时候审计和动态实时的风控。通过异常检测算法对站点风险进行消解。
- 运维助手：快速帮助运维人员获取故障信息和故障恢复方案长期是一个难点，高质量的运维需要运维人员有丰富的相关经验。传统方式需要运维人员查询大量的文档来制定和完成运维工作。随着生成式AI的发展，通过LLM+垂域知识的RAG方案实现运维领域的智能问答系统，来辅助运维人员来完整运维工作。运维人员通过文本语音和助手进行交互，助手基于用户的意图提供运维内容检索，故障根因推荐，信息查询，故障报告生成等能力。

随着生成式AI能力高速发展，结合AI Native的概念深入，运维能力的建设从辅助运维逐步发展成以数据和大模型为中心的主动运维能力。当前的运维大脑采用的是多AI Agent的架构如下图：

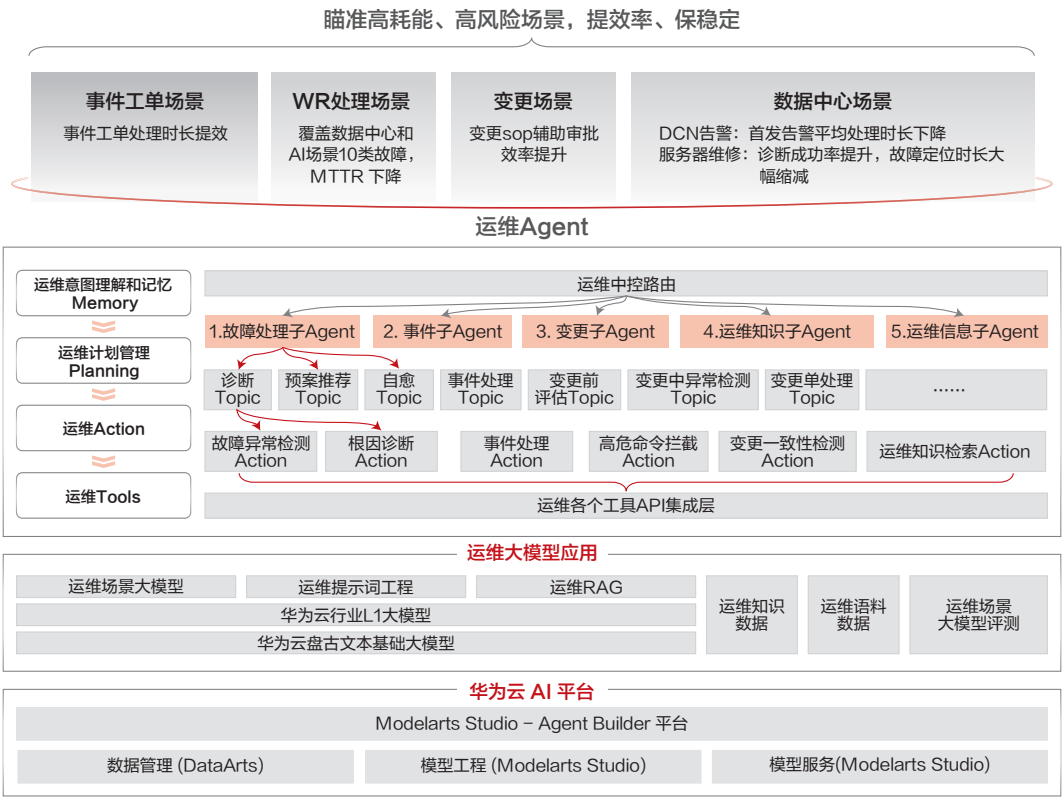


图108 AI Native 多Agent架构运维大脑

对于事件工单场景，WR处理场景，变更场景，数据中心场景等对外提供了一个统一的运维Agent，同时运维Agent又包含了5个子Agent：故障处理Agent、事件Agent、变更Agent、运维知识Agent、运维信息Agent。这些Agent都是基于Modelarts Stuido的 Agent开发套件构建。 运维Agent通过环境感知，共享长短记忆，基于推理大模型实现运维规划，大模型执行规划和结合大模型的工具应用，来协同五个子Agent来完成运维分析和运维计策。 运维人员只需要向Agent提交运维问题，Agent会基于用户的问题自动的把相关的知识、诊断、方案、决策输出。 通过AI Agent实现了搞耗能、高风险场景的，提效率和保稳定的目标。

3.6 AI-Native应用

3.6.1 AI Agent成为确定性未来

大模型在基础能力上快速突破拐点，尤其是在长任务推理、工具调用等方面，使得AI Agent快速从“同步对话”、走向“异步任务”、甚至“自主智能”，AI Agent也成为调用传统软件、工具的智能入口；从企业视角看，以AI Agent实现对生产流程、生产工具的深度重塑、提升生产效率，已成为一种确定性的未来。2025年初，随着Manus、Genspark等通用AI Agent的一度爆火，以及MCP（Model Context Protocol）生态的蓬勃发展，更是进一步推动AI Agent应用在行业的快速发展。

从企业视角，可以分三个层次来理解为什么AI Agent会成为一种必然的趋势。首先，大模型作为人类“知识与经验”的压缩，可以有效承载持续累积的领域知识与经验，并通过模型的再生成能力，以近乎零成本实现“知识与经验”的传承及复制。

其次，基于AI大模型及领域知识库能力，可以构建各种丰富的场景Agentic AI技能，赋能从方案设计、产品开发、营销方案，到客户服务等诸多业务领域，以一种类人、或者超人的普适性实现超越普通技能、甚至熟练技能的智能效果。

最后，大模型及AI Agent可以说从根本上重塑了人机交互模式、内容生产方式、以及软件定制方式，从而使得AI Agent成为普适性的“工具的革命”，驱动各行各业生产效率的跃升。

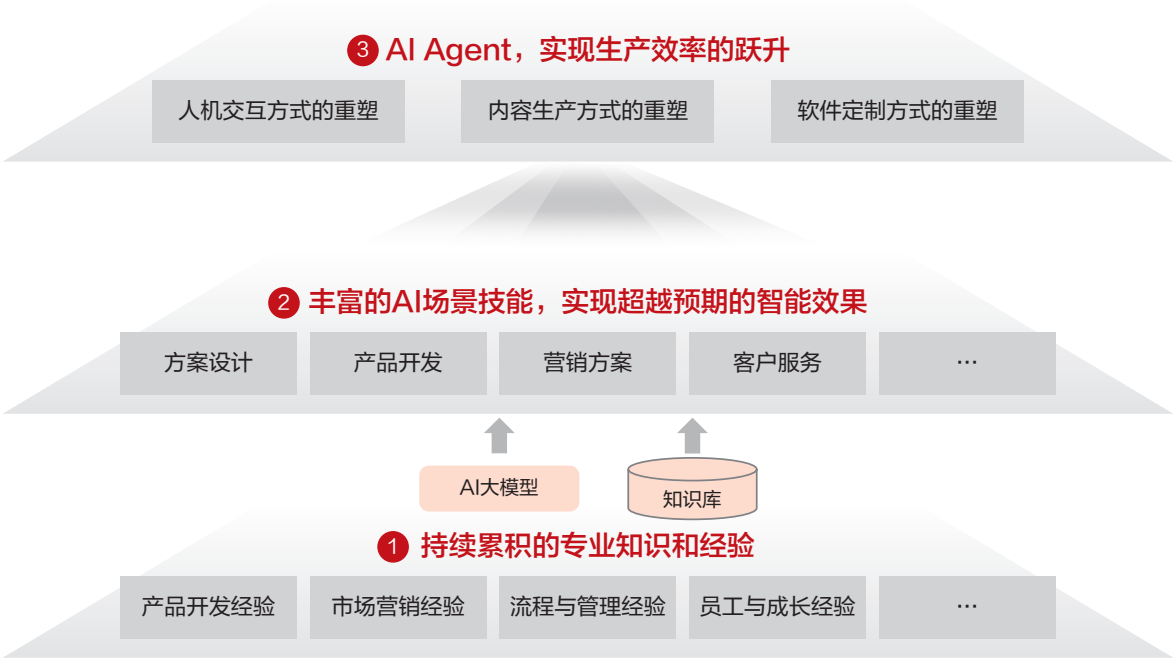


图109 企业AI Agent场景及价值层次图

所以，无论从技术视角、还是业务视角，AI Agent不仅成为一种必然、而且当前具备了落地的技术可行性。从企业实践场景，围绕人人都需要的“员工助手Agent”、到各类专业领域Agent，可以形成一张AI Agent协同的网络。

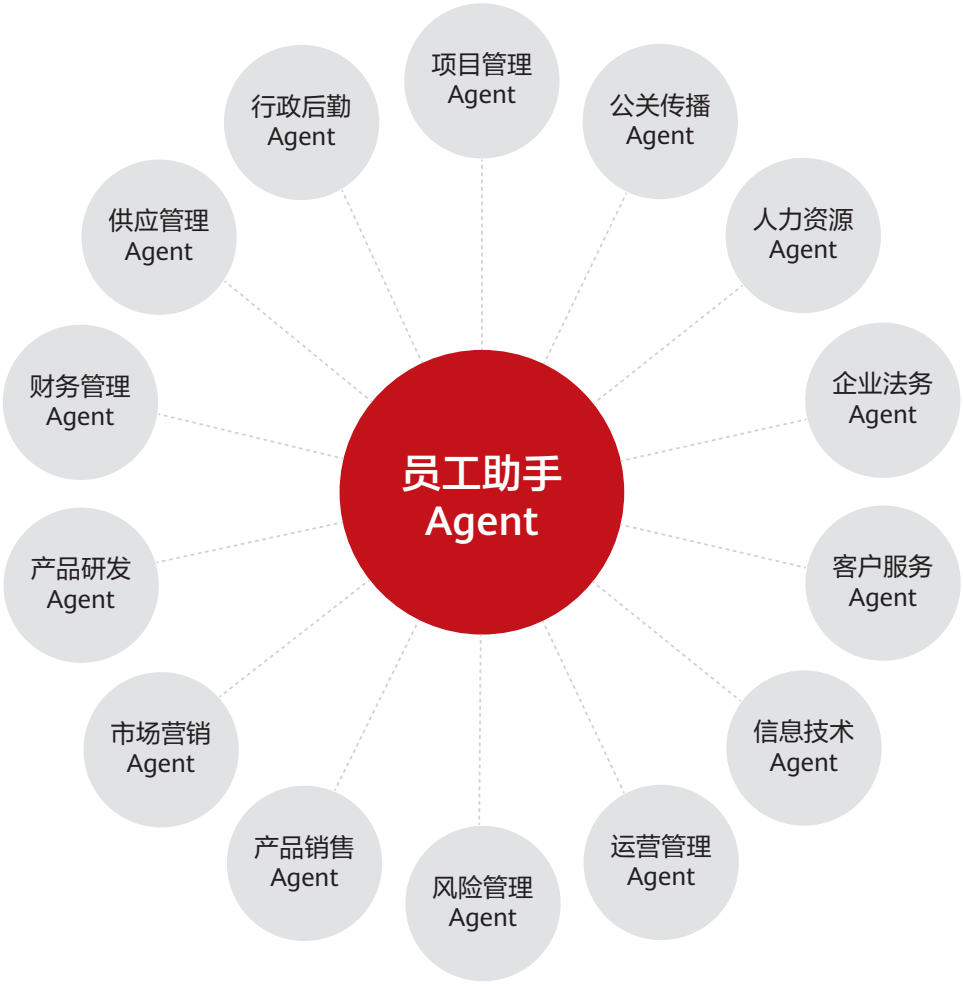


图110 企业AI Agent协同网络

3.6.2 AI Agent与AI-Native应用架构

大规模及AI Agent代表软件发展历程上的一个范式转折点。AI Agent系统是大模型动态地编排业务流程和调用工具使用的智能系统。这代表了一种由LLM驱动的编排形式，其中模型的智能被用于规划、推理并决定下一步行动。AI Agent不是遵循预定义的脚本，而是在一个循环中运行，接收一个高层次的目标，推理如何实现该目标，选择并使用工具，观察环境中的结果，然后决定下一步行动。传统软件是基于预定义代码做出决策，而AI Agent则是基于模型本身生成的实时预测和推理做出决策。这赋予了AI Agent一定程度的自主性，使其能够根据环境上下文调整策略，并处理那些无法预先硬编码的逻辑及其顺序的开放性问题。

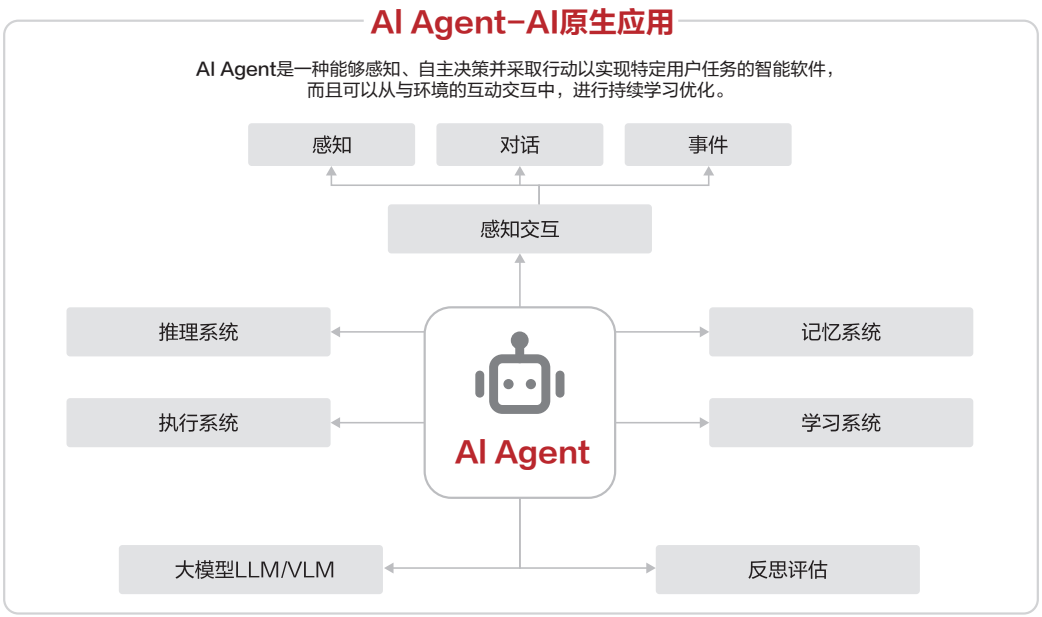


图111 基于AI Agent的AI原生应用

这种动态的、由模型驱动的决策方式使AI Agent在处理具有模糊性和不可预测性的任务时表现出非凡的能力。它们在解决方案路径事先未知的场景中表现出色，例如复杂的研究、动态的客户支持对话或多步骤的故障排除。AI Agent可以探索不同的解决途径，通过尝试替代方法来从错误中恢复，并从单次、多次交互中的反馈中学习。然而，这种能力也带来了显著的复杂性、运营成本和可预测性方面的代价。AI Agent可以取得显著的成果，但也可能表现出难以预料、难以调试，甚至可能代价高昂的行为。

3.6.3 AI Agent应用开发框架

对于AI Agent应用来说，需要一个与之匹配的开发框架及开发工具链，总体上可以分为“AI应用工具链”、及“AI应用运行时”两大部分。

“AI应用工具链”一般包括面向AI应用的构建工具，也就是AI应用的Dev工具部分，在构建方式上分为：

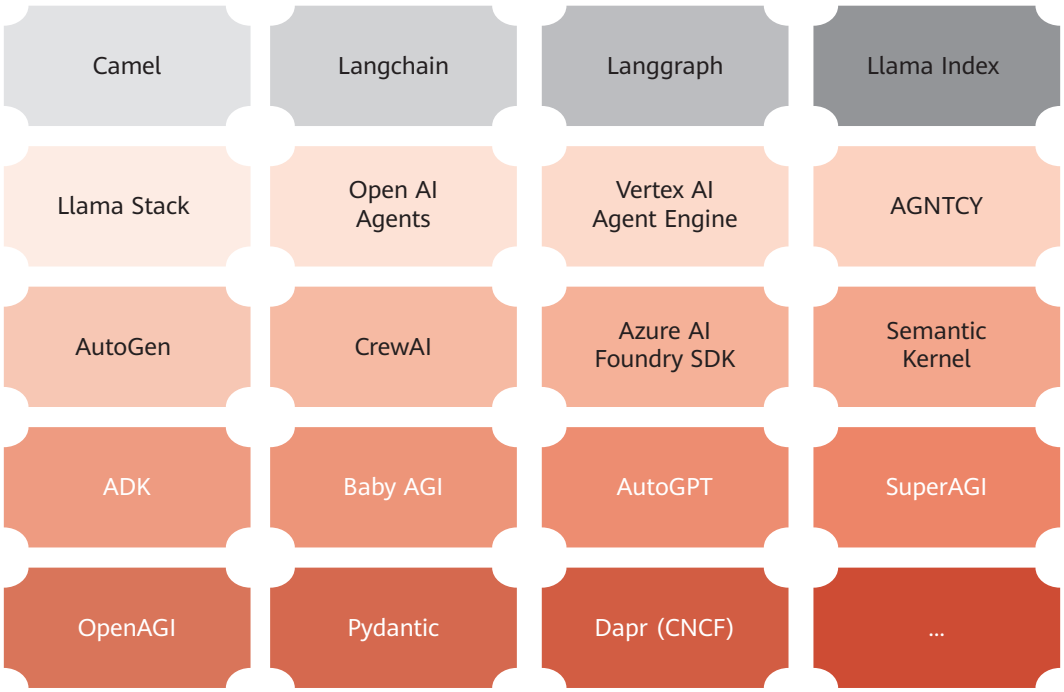
- 零码构建：比较常见的“一句话”生成AI Agent，通过Agent框架串接大模型能力，自动补齐知识库集成、及工具调用，从而实现一个场景AI Agent；也支持用户对系统提示词、知识库、工具的配置进行修改。
- 低码构建：比较常见低码工作流编排模式，通常用业务SOP或人工拆解完成业务流程的定义、模型工具、模型知识上下文、以及流程逻辑控制的实现，适合确定性比较高的场景。
- 高码构建：基于流行的Agent SDK通过Python等高级编程语言，用传统的Python等工具链开发的一种AI应用开发模式。

图112 企业AI Agent全栈参考架构

除构建工具外，当前也出现了许多面向AI原生的各类Ops工具，为AI应用提供监控、评估及优化这三大能力。

“AI应用运行时”一般可以分为四层，分别为：数据服务，也就是常见的知识库、RAG、Web搜索、MCP协议等部分；模型服务，也就是常见的模型网关负责模型请求路由、内容生成合规过滤、以及模型与数据安全等；规划与编排，一般负责意图路由、业务流程的自主规划或者基于工作流编排，以及应用的上下文记忆能力；AI模型应用，目前主要分为智能体应用（AI Agent）、以及非智能体应用（各种AI使能的应用）两种。

目前对于AI Agent类应用的开发，出现了一批开源的Agent Framework（框架），极大的便利了Agent应用开发的便捷性。



一部分开源Agent开发框架举例

图113 业界开源Agent开发框架

3.6.4 华为云Versatile Agent平台

华为云Versatile面向AI Agent应用，提供了完备的AI-Native应用运行时服务及丰富的AI-Native的Dev、Ops工具能力，可以赋能个人开发者、企业开发者，实现基于大模型的AI应用快速创新。

华为云Versatile的构成如下：

图114 华为云Versatile Agent平台

AgentStudio:

- 定位为AI应用工程，Agent的开发工具
- 打造全开放，强运营的AI应用使能平台

AgentSpace:

- 为用户提供统一Agent入口，Agent统一交互入口，人与AI协同的智能入口，统一的Agent指挥中心
- Versatile AI Agent，面向企业的通用AI智能体，官方智能体+伙伴智能体

AgentGallery:

- 提供Agent/MCP/Tools 接入能力
- 提供全套工具链及托管环境，在开发态由客户选择AI产品，实现动态集成

AgentBase:

- AI亲和的文档数据底座和工业数据底座
- 提供数据建模、知识管理、上下文记忆
- 自演进强化学习框架

AgentRun:

- AI原生应用运行时服务，支持三方Agent应用接入，托管
- 虚拟机级资源隔离；实例基于微虚拟机（microVM）极致弹性

AgentOps:

- 承载Agent从“构成”到“养成”使命
- 为Agent应用提供端到端完整的追踪、测试、部署和监控，以及根因分析与自适应优化功能

3.6.5 华为云AI Agent应用工程实践

AI Agent在企业的落地，从本质上看是一场技术与业务融合的变革，不仅需要自顶向下的顶层设计与业务规划，也需要自底向上的全员参与及实践推广。从业务、技术两个维度面临许多新的挑战，归纳如下：

表8 AI Agent企业落地挑战

	挑战	描述
1	如何选择合适的大模型	当前模型的种类很多，且能力各有所长，如何对模型进行有效评测，选择匹配业务场景的最佳模型；另外模型发展迅速，如何构建模型引入、管理的标准和机制，对模型进行有效的治理，避免潜在的混乱和风险；
2	如何发挥大模型的能力	大模型本身是一种无状态的、标准化产品，如何为大模型构建业务场景提示词、如何编排不同的模型，以及如何基于业务数据回流，提升模型的能力等；
3	如何准备高质量数据	高质量数据决定智能的高度，如何准备高质量的行业知识、企业私域数据集，如何有效的向模型和应用供给上下文数据等；
4	业务与技术部门如何协同	AI 应用的构建、运营依赖 IT 部门和业务部门的充分协同和配合，它与传统应用的开发工程和开发流程有很大区别，如何重新梳理 IT 与业务部门之间，以及 IT 部门内角色间的职责与协同关系等；
5	如何确保安全与合规	数据是企业高价值资产，如何确保与模型交互的数据安全、或者用私域数据增量训练、微调后的模型文件安全，以及保护与外部 MaaS 服务交互时的数据安全及隐私安全，都是非常大的技术挑战；此外，智能体应用在提示词、意图路由、幻觉抑制，以及身份与权限隔离等领域，还带来许多新的挑战；
6	如何平衡成本与收益	如何统筹规划算力资源、模型资产、数据资产等，以实现集约化、高效利用；以及在运营层面制定标准、方法和工具，持续观测、以及优化成本与收益；

为了有效的解决这些新的挑战，以及保障AI Agent落地的成功率及业务效果，华为以“智能化变革”为牵引，并根据内部各业务领域的实践总结了一套企业进行AI Agent落地的系统化方法，也称为“五阶八步”的AI Agent落地最佳实践：

图115 企业Agent落地“五步八步”法

“五阶八步”特指五个阶段八个关键任务，分别包含：场景识别、流程重构、组织能力准备、数据与知识准备、以及AI Agent开发与AI Agent持续运营，从过程上对AI Agent落地给出了切实可行的指导方法。这其中，最具挑战性的是“场景识别”及“目标设定”，因为它不仅依赖大模型的能力成熟度、也与企业侧的数据、知识的准备度，以及效果期望紧密相关。

为了更好的指导“场景识别”与“目标设定”，华为在实践的基础上细化了“AI场景十二问”，分别从“商业价值”、“场景成熟度”、“持续运营”三个板块，进行度量、筛选。下图是“AI场景十二问”具体实践时的操作参考。

AI场景“十二问”			说明：填写场景名称，可按具体场景，或场景分类分析
D1：商业价值		①业务场景是否能清晰度量价值	说明：展示落地该场景的价值点
		②落地后收益评估，3年内ROI是否为正	说明：当满足，可打“√”；不满足，可打“×”；不涉及该项的话，填写“不涉及”
D2：场景成熟度	业务准备度	③业务场景有明确的业务Owner（对投资和结果负责）	说明：
		④业务场景有明确的流程规则（业务说得清）	说明：当具备，可打“√”；不满足，可打“×”；部分明确该项的话，填写“部分明确”
		⑤业务场景有明确的用户触点（业务已数字化）	说明：当具备，可打“√”；不满足，可打“×”；部分明确该项的话，填写“部分明确”
	数据准备度	⑥业务知识/数据是否足够支撑0~1冷启动（范围清晰、完整、易获得）	说明：当具备，可打“√”；不满足，可打“×”；部分明确该项的话，填写“部分明确”
		⑦业务知识/数据是否随作业持续产生、更新和反馈	说明：当是，可打“√”；不是，可打“×”；
	技术准备度	⑧现有技术能力是否能够支撑场景实现（技术可行性）	说明：当是，可打“√”；不是，可打“×”；
		⑨云内部、公司内是否有成功经验可以复用/借鉴	说明：当是，可打“√”；不是，可打“×”；
D3：持续运营		⑩有清晰的业务运营目标	说明：若有，正常填写；若没有，填写“需要明确”
		⑪业务目标有运营数据支撑（过程可度量）	说明：若有，可打“√”；若没有，可打“×”；云服务Agent该项均有，可打“√”
		⑫业务有持续运营的组织、资源、机制和能力	说明：若有，可打“√”；若没有，可打“×”；云服务Agent该项均有，可打“√”

图116 华为实践AI Agent场景筛选与评估方法

从华为自身的AI Agent实践来说，始终将AI的可信与安全作为最高优先级，这也代表企业类用户、以及生产类场景在落地AI应用时所必备的治理能力、及平台能力。从AI的安全可信管理总纲、及全过程管理来说，可以总结为下图的内容：

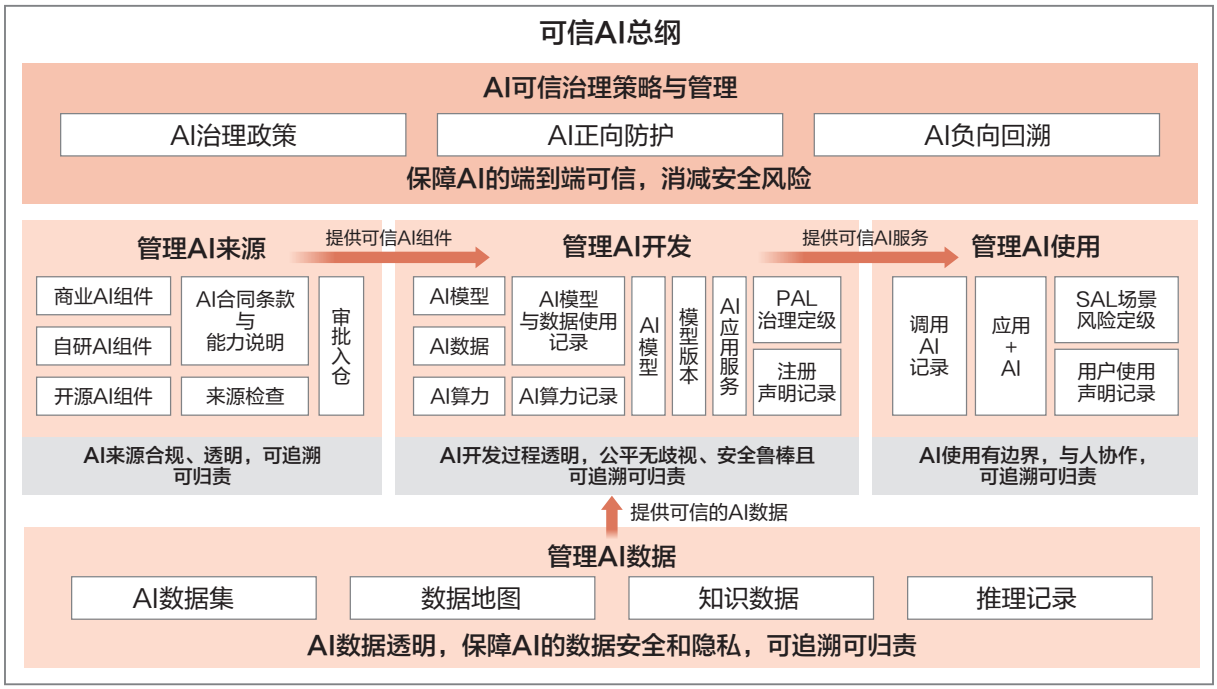


图117 华为实践AI应用的可信治理与安全防护措施

AI Agent及其技术栈的构建和演进，离不开一个成熟的生态体系。华为云依托全球开发者与生态伙伴力量，构建了完善的AI技术生态体系，并持续贡献各类高价值AI资产及AI能力，可以持续赋能行业应用的智能化重塑以及行业AI Agent的创新。

3.7 大模型安全

大模型在蓬勃发展的同时，自身也面临着严峻的安全和合规问题，包括数据泄露和内容安全的问题、大模型自身遭受攻击、日益严峻的监管要求等等。

3.7.1 大模型面临的安全风险与合规要求

» 3.7.1.1 大模型应用面临的安全风险

大模型应用快速发展，应用场景和领域在不断扩大。基于大模型的应用提供越来越多的能力。在这个过程中，出现了大模型提示泄露、大模型越狱、大模型Agent权限管理等多种安全漏洞。MITRE ATLAS（Adversarial Threat Landscape for Artificial-Intelligence Systems）对人工智能应用进行威胁分析，当前给出了14种威胁战术，以及这些战术中用到的67种威胁技术。

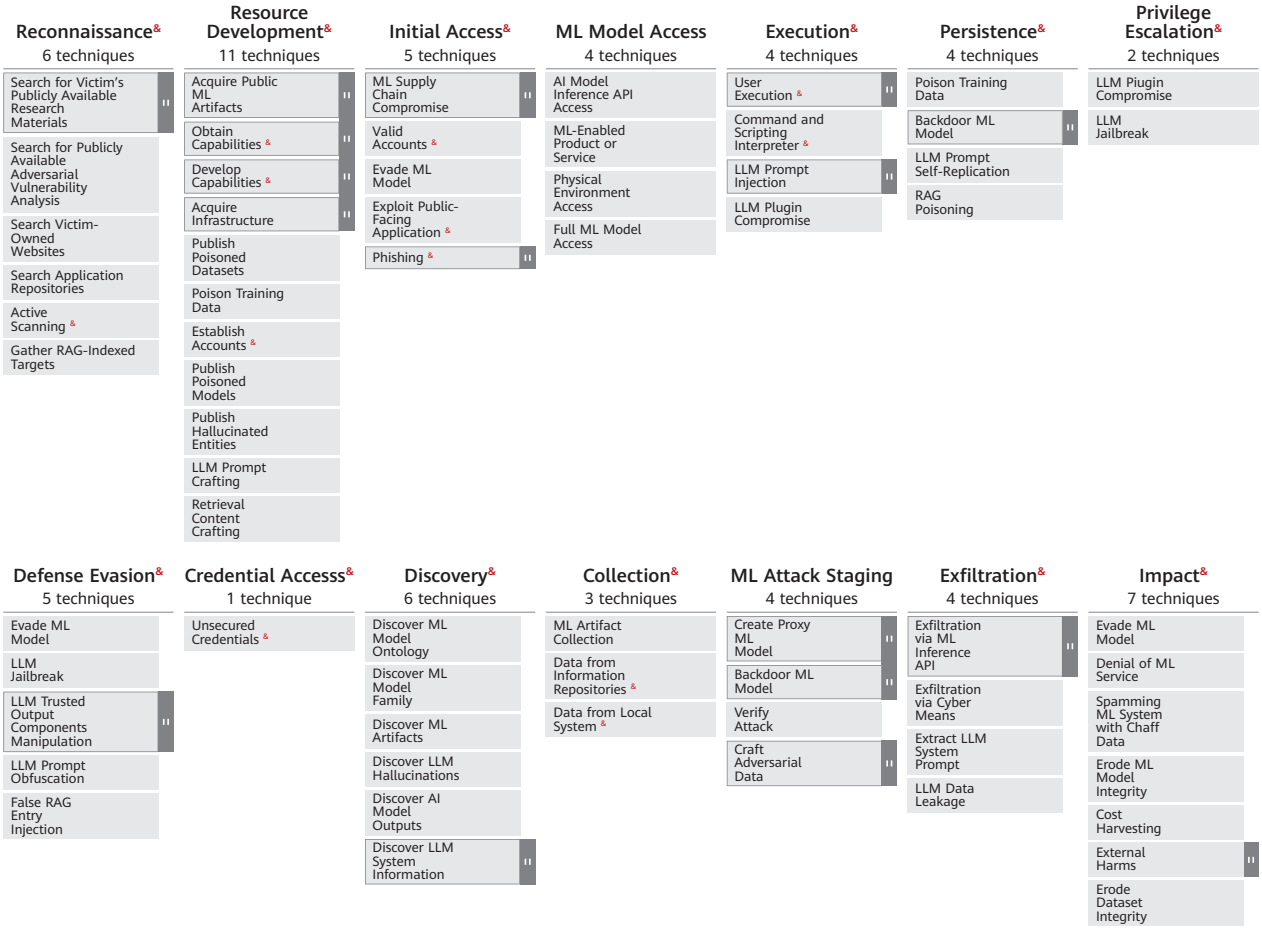


图118 MITRE ATLAS人工智能系统的对抗性威胁全景

对于这些威胁带来的风险，OWASP在2025年给出了大模型应用的TOP 10安全风险。

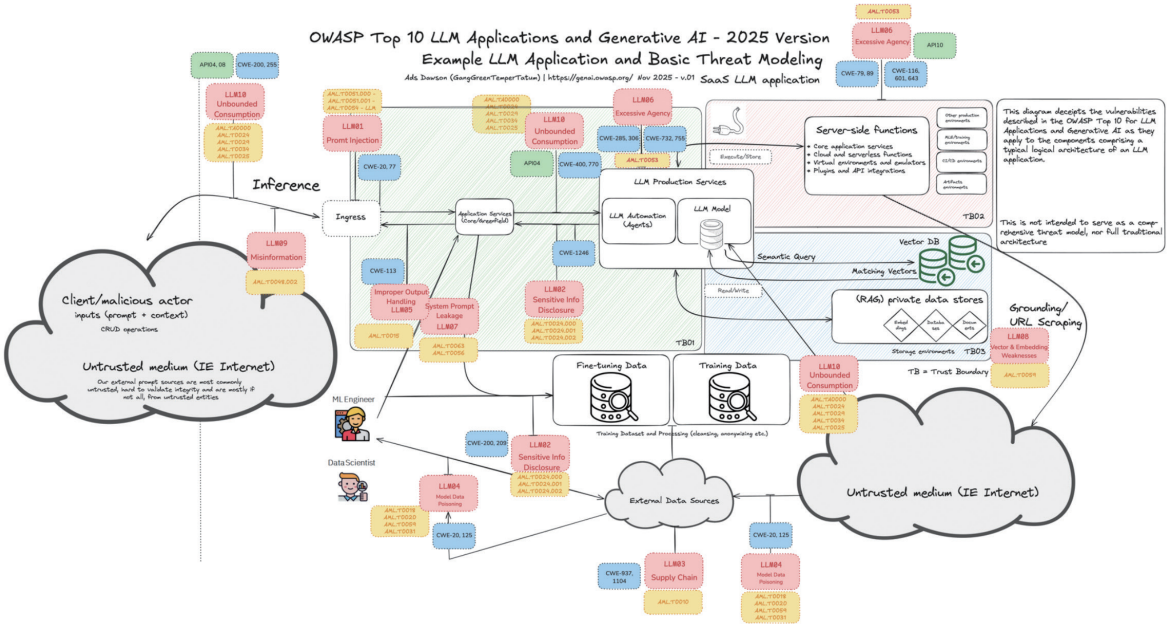


图119 OWASP大语言模型与生成式AI十大风险（2025）

风险包括与语料数据安全相关的LLM04数据中毒、LLM02敏感信息泄露、LLM08向量嵌入不安全、LLM03不安全数据供应链风险等风险。与模型相关的LLM03模型供应链风险、LLM06代理权限过大、LLM01提示注入攻击、LLM10无限制资源消耗攻击、LLM07系统提示泄露风险、LLM05不当输出处理、LLM09虚假与误导信息、LLM02敏感信息泄露等风险。

- LLM01提示注入风险：当用户以非预期的方式提示 LLM 改变其行为或输出时，会发生提示注入攻击。提示注入涉及通过特定输入操纵模型响应以改变其行为，这可能包括绕过安全措施。越狱是一种特殊的提示注入，攻击者提供输入，导致模型完全无视其已有的安全协议。
- LLM02敏感信息泄露：基于大模型提供的应用，存在泄露敏感与隐私数据、专有算法与机密信息等风险。可能导致未授权的数据访问、侵犯隐私和知识产权泄露。
- LLM03供应链风险：大模型的供应链容易受到多种漏洞的影响，这些漏洞可能会损害训练数据、模型和部署平台的完整性。这些风险可能导致输出偏差、安全漏洞或系统故障。在传统软件漏洞中，通常关注代码缺陷和依赖关系，但在机器学习（ML）领域，风险还扩展到了第三方预训练模型和数据。
- LLM04数据与模型投毒：数据投毒是指在预训练、微调或嵌入数据中被人为操纵，目的是植入漏洞、后门或偏见。这种行为可能会破坏模型的安全性、性能或道德标准，造成有害的输出或性能下降。常见的风险包括降低模型效能、产生偏见或有害内容，以及对下游系统的潜在威胁。
- LLM05不当输出处理：不当输出处理是指在大模型生成的输出被传递到其他组件和系统之前，没有进行充分的验证、清洗和处理。
- LLM06过度代理：基于大模型的系统通常被开发者赋予一定程度的代理权限，即它们能够通过扩展调用函数或与其他系统交互，以响应提示并执行操作。过度代理是一种安全漏洞，它依据大模型产生的意外、模糊或被操纵的输出做出响应，执行有害的操作。过度代理的根本原因通常涉及一个或多个方面功能过度、权限过大、自主权过高等。

- LLM07系统提示泄露：LLM中的系统提示泄露漏洞指的是用于指导模型行为的系统提示或指令可能暗含未被察觉的敏感信息。这些系统提示虽然旨在引导模型根据应用需求输出结果，却可能无意中夹带了秘密信息。一旦这些信息外泄，就可能被用来发起其他攻击。
- LLM08向量和嵌入不安全风险：在采用检索增强生成（RAG）技术的大模型系统中，向量和嵌入的漏洞构成了重大的安全风险。在向量和嵌入的生成、存储或检索过程中，存在的弱点可能被恶意行为者（无论是有意还是无意）所利用，导致有害内容的注入、模型输出的操控或敏感信息的泄露。
- LLM09虚假与误导信息：大模型制造的虚假信息给依赖这些模型的应用程序带来了核心风险。所谓虚假信息，指的是 LLM 生成的内容表面上看起来很真实，但实际上却是错误的或具有误导性的。这种漏洞可能引发安全漏洞、声誉受损以及法律责任等问题。
- LLM10无限制资源消耗漏洞：无限制资源消耗风险指的是大型语言模型（LLM）根据输入的查询或提示来生成输出的过程。推理是LLM的核心功能之一，它涉及到运用已学习的模式和知识来产生相关的回答或预测。无限制资源消耗的目标是破坏服务、耗尽目标的财务资源，甚至通过模仿模型行为来窃取知识产权的攻击，都依赖于此漏洞。

多模态大模型融合了文本、图像、音频、视频等多种模态能力，相对于语言模型有结合其他模态独特的攻击技术。如对多模态大模型的提示注入攻击，采用良性文本和恶意图像结合的攻击方式。对抗样本与中毒样本的形式相对于语言模型也有区别，如图像噪声、音频干扰、特定图像模式的触发器。多模态大模型带来更大的虚假内容风险，逼真的虚假图像、音频、视频带来的隐私与版权侵犯、误导与偏见等。

» 3.7.1.2 大模型安全相关的法规与标准

对于人工智能应用引入的安全风险，部分国家和国际组织制定了相关的法律法规和标准。2023年8月15日，中华人民共和国七部委联合发布了《生成式人工智能服务管理暂行办法》，2024年2月29日，中华人民共和国网络安全标准化技术委员会发布了《生成式人工智能服务安全基本要求》技术标准。2024年8月1日欧盟生效了《人工智能法案》。2024年7月26日，美国NIST发布NIST-AI-600-1《人工智能风险管理框架：生成式人工智能概况》；国际标准组织ISO/IEC发布了《ISO/IEC TR 24028:2020信息技术-人工智能-人工智能中的可信性概述》等系列技术规范。

中华人民共和国《生成式人工智能服务管理暂行办法》在总则提出了坚持发展和安全并重、促进创新和依法治理相结合的原则，进行了价值观、知识产权、隐私保护、可信性等方面要求。在技术发展与治理章节鼓励人工智能技术应用与相关技术创新合作，进行了训练数据处理活动方面的规定；在服务规范章节对提供者进行了内容生产责任等方面的相关规范要求；在监督检查和法律责任章节要求政府相关部门依法加强对生成式人工智能服务进行管理。《生成式人工智能服务安全基本要求》技术标准给出了语料来源安全、语料内容安全、语料标注安全、模型安全、模型生成内容安全等方面要求。

欧盟《人工智能法案》（AIA）根据人工智能系统对个人基本权利、健康或安全，以及整个社会的潜在风险，将AI系统分为四个风险等级，每个等级对应不同的监管要求。高风险AI系统必须进行风险管理、数据治理和管理实践，保障技术文档、日志能力、透明度、人类监督、准确性、鲁棒性和网络安全。

美国NIST《人工智能风险管理框架》包括四大功能，即治理（govern）、映射（map）、测量（measure）和管理（manage）。每个功能项下还分为不同的类别和子类别。治理功能包括政策、流程、问责等六个类别；映射功能包括系统分类、系统功能、系统影响等五个类别；测量功能包括评估系统可信性、完善风险识别跟踪机制等四个类别；管理功能包括对系统风险进行判定、响应，明确风险响应步骤等四个类别。

《ISO/IEC TR 24028:2020信息技术-人工智能-人工智能中的可信性概述》中给出了AI面临的威胁与消减措施。威胁包括数据投毒、对抗攻击、模型窃取等AI安全威胁，数据获取、处理数据、模型查询等AI隐私威胁，偏见、不可预测性、不透明性、硬件故障等其他威胁。消减措施包括透明度、可解释、可控、消除偏见、隐私保护、可靠韧性健壮性、Safety、评估与测试等。

» 3.7.1.3 大模型面临的风险消减与合规需求总结

针对大模型面临的安全风险与合规需求，需要进行综合性安全体系建设，来消减上述各种安全风险，以及保障合规遵从。华为云依据华为30年以上安全建设经验，总结了一个中心、七层防线安全防护框架，包括物理安全、账号权限安全、网络安全、应用安全、主机安全、数据安全、运维安全以及统一安全运营中心。大模型是网络环境中的一种类型的应用，与大模型密切相关的防范在应用防线和数据防线，以及对模型的安全运营。

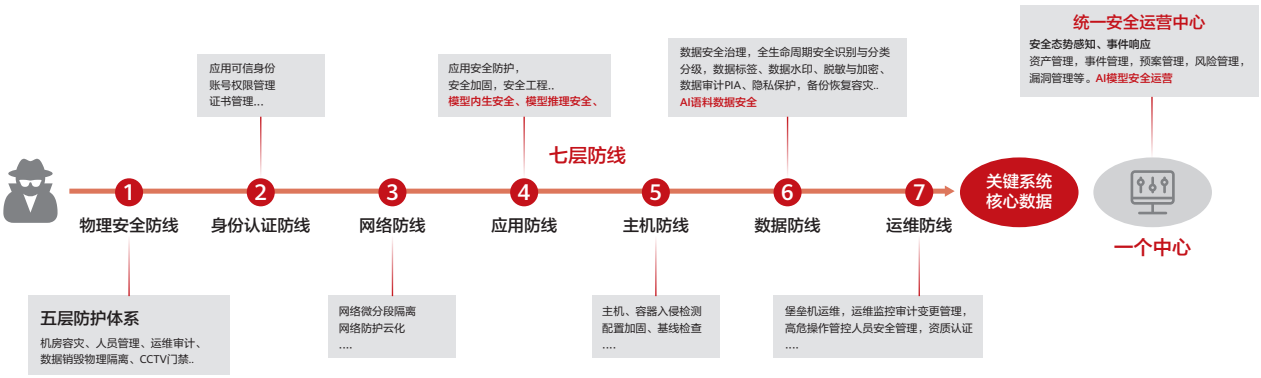


图120 华为云一个中心、七层防线的安全防护框架

在数据准备与数据标注阶段需要进行数据来源可信、数据内容合规、数据版权合规、数据中毒与后门检测、敏感与隐私保护、知识库安全、数据安全防护、数据防篡改、数据访问控制、数据加密等安全防护。

在模型开发阶段保障模型内生安全，需要进行包括数据、模型、软件的供应链安全、对抗鲁棒性训练、价值观对齐训练、模型安全测评、模型防篡改、模型访问控制、模型加密等安全防护。

在模型推理阶段需要进行提示注入攻击防护、主题保持与幻觉防护、生成内容安全合规、资源滥用防护、生成内容追溯、生成内容安全使用、智能体安全、模型防篡改、模型访问控制、模型加密等安全防护。

对大模型进行的安全运营，需要进行模型资产风险预防、模型安全态势感知与事件响应、运行环境安全运营、用户申诉处理等安全运营工作。

对于大模型运行的环境，需要进行物理安全、账号权限安全、网络安全、应用安全、主机安全、数据安全、运维安全等安全防护。

3.7.2 大模型安全技术

» 3.7.2.1 数据安全

- 数据内容合规：为保证大模型生成内容合规，首先要对语料数据的合规性与安全性进行防护，相关技术有预训练语言模型（如BERT、RoBERTa）构建细粒度文本分类器，通过海量标注数据训练实现敏感语义特征提取，覆盖涉黄、赌博、诈骗等典型违规场景，以及结合知识图谱进行检测等。
- 数据版权合规：在使用任何数据集前，应仔细检查并确认其许可协议。遵循如Creative Commons、Apache License等开放许可协议的数据集。在使用开源数据集时，选择标记为开源和用于公共领域的数据集。第三方数据集时数据根据数据许可协议，采用合规的数据集。建立数据集版权库，明确各类数据集的来源和适用的许可协议。针对公开获取的数据，通过版权扫描工具自动进行数据成分分析，识别具有版权的内容，检测是否有侵权风险。对于向外提供的自有数据需要进行水印防护，防止未经授权用于大模型训练。相关技术有文档水印、图片水印、结构化数据水印等。文档与图片支持明水印与暗水印，数据水印一般是不可见水印。
- 敏感信息检测与隐私保护：利用自然语言处理（NLP）、机器学习或深度学习技术自动识别文本中的敏感信息，如身份证号、电话号码、地址等。进行多维度识别，不仅要识别标准格式的敏感信息，还要能识别变体、模糊化表达或隐晦提及的敏感内容。检测识别需要支持上下文理解：在识别过程中考虑上下文，避免误判，比如区分真实的个人信息与文学作品中的虚构信息。数据脱敏方法有替换、掩码、泛化、混淆、删除等，以适应不同类型的敏感信息和应用场景。脱敏技术有可逆/不可逆脱敏：根据需要选择是否保留数据恢复的可能性。例如，使用哈希或加密技术实现不可逆脱敏，或使用假名化技术实现可追踪的脱敏。脱敏需要对模型兼容，脱敏技术需能应用于不同类型的大模型，包括但不限于生成式和判别式模型，且不影响原来模型的训练和推理效果。在训练数据有特殊隐私保护需求的场景，采用差分隐私、联邦学习、多方安全计算等技术在保证数据隐私的情况下，进行训练相关的计算。
- 知识库安全：组织数据在敏感程度与知悉范围上存在安全要求，需要基于角色进行自有知识安全隔离。可以采用基于多知识库、基于标签、基于向量数据库Metadata的方式进行过滤。
- 数据安全防护：数据安全防护主要包括用户数据防范技术、依赖数据库的安全防护技术，包括数据库漏扫、数据库加密、数据库防火墙、数据脱敏、数据库安全审计系统等。
- 数据中毒与后门检测：训练数据中如果有毒性样本或者带有后门的样本，会导致模型在推理阶段产生错误内容和不安全内容，采用对抗算法、可疑触发单词检测等技术发现相关样本，进行去除相关处理。

» 3.7.2.2 模型内生安全

- 对抗鲁棒性训练：大模型鲁棒性是指的面对输入数据的变化、噪声和攻击时，仍然能够保持稳定的能力，例如语句变化、开放性问题、不可理解内容、图像变化、声音噪声等。采用数据增强和对抗训练，可以提升此类问题的表现。
- 价值观对齐训练：监督微调（Supervised Fine-Tuning，SFT） 是一种通过标注数据对大模型进行针对性调整的技术，使用人工标注或者精心设计的指令数据集对大模型进行监督微调，可以使模型更好的回答问题。人类反馈强化训练（Reinforcement Learning from Human Feedback，RLHF） 是一种通过结合人类反馈和强化学习（RL）来优化大模型的技术，让模型的输出更符合人类偏好，如真诚、善良等。

- 供应链安全：在大模型开发过程采用可信开发过程实践，对大模型相关的数据、模型、开源以及三方件供应链，进行选型、防篡改、漏洞修补等进行全生命周期安全管理，消减供应链安全风险，如建立AI相关的BOM等。可信开发过程详细内容可参考《华为云可信白皮书》。

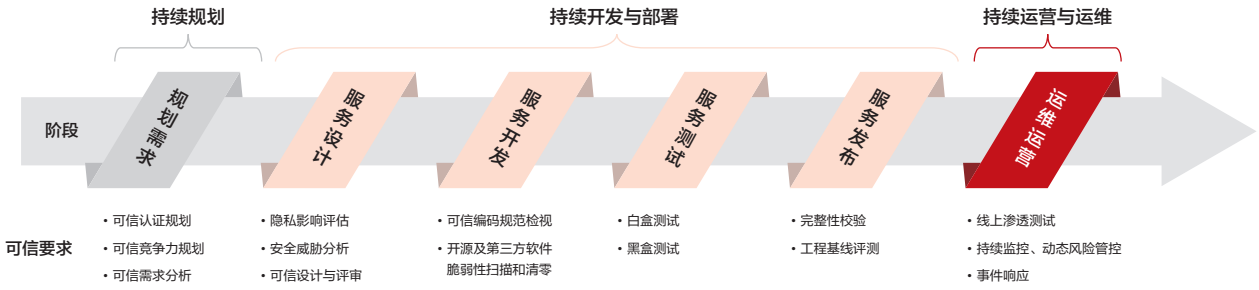


图121 华为云可信开发过程实践

- 模型安全测评：通过系统化的方法评估大语言模型在安全性、可靠性、伦理合规性等方面的表现，包括自动化测试、人工测试、大模型红队测试等方式对模型能力进行验证。

» 3.7.2.3 模型推理安全

- 提示注入攻击防护：基于攻击与正常数据，预训练Prompt攻击检测的Bert模型；对于开放域的Prompt攻击，预训练模型不能解决所有问题，可使用敏感词匹配，与向量检索能力进行综合决策构建完整可靠的Prompt攻击检测能力；对多种攻击手法（例：目标劫持、角色扮演、反面诱导、悖论攻击、初始肯定等）进行检测，对单轮和多轮交互攻击手法进行检测。
- 主题保持：在生成阶段对生成内容进行约束和后处理，保障生成内容检测是否符合预期主题、避免偏离或敏感信息泄露。相关方法有收集整理行业语料，使用语义模型进行训练；提取输入提示词与生成文本的语义嵌入，计算余弦相似度或欧氏距离，判断是否偏离阈值。
- 幻觉防护：大模型存在生成与输入源不一致内容的情况，称为大模型幻觉，原因可能是训练数据中源和目标存在偏差情况、训练数据有重复、数据噪声等。缓解幻觉的方法有利用外部知识验证，如结合知识图谱，采用RAG，加强大模型数据生成质量。
- 生成内容安全合规：对生成内容进行审核，包括不仅限于以下方面：个人隐私、伦理道德、偏见歧视、军事 、商品推荐、宗教信仰、心理健康、政治、文化、暴恐、民族、网络攻击、自身、色情、财产隐私、赌博、身体伤害、辱骂、违禁。基于分词的滑动窗口技术：将大模型生成的token进行分词并基于分词进行滑窗，确保每次仅对必要的数据进行匹配。内容审核通过Bert、Fasttext、HMM、Word2Vec等一系列NLP技术，以及声音与图像相关的LSTM、CNN等神经网络技术。对生成内容进行敏感信息与隐私信息检测，进行相关的脱敏处理。
- 资源滥用防护：一些用于模型窃取的调用会消耗模型推理资源，如通过大量输入的输出得到大量数据训练本地模型，用一些对抗性输入触发模型暴露更多内部逻辑。对于此类攻击，可以通过限制API调用频率、对异常输入进行拦截等技术进行防范。

- 生成内容安全使用：大模型生成的内容不能完全可靠，需要将风险告知用户。使用大模型的输出需要进行必要的处理，如大模型生成的代码、payload可能造成安全漏洞风险，建立应用程序安全开发验证相关的规范、流程、验证工具，可以消减此方面风险。基于大模型构建的智能体应用，会依据大模型的输出与其他系统交互，大模型输出内容不可靠会造成应用安全风险。此种情况需要依据系统交互做威胁分析，对Agent的权限、对系统的调用做最小化控制。
- 生成内容追溯：对大模型生成的内容增加水印有助于对内容进行追溯，对于不同模态的生成内容采用不同的加水印技术。
- 智能体安全：基于大模型的智能体面临任务感知、任务编排、任务执行等风险，需要加强任务意图理解、最小权限任务编排、最小权限API的调用等方法消减智能体风险。

» 3.7.2.4 安全运营

为了保障模型安全运行，需要进行大模型相关全范围的安全运营，在账号权限、网络、主机、运维、数据、应用等各方面进行风险预防与态势感知，对于发现的威胁进行事件分析与响应处理。下面是与大模型密切相关的几个安全运营方面。

- 模型资产风险预防：建立模型资产清单，对资产进行安全漏洞扫描、安全基线检查以及其他安全评估验证。如对模型依赖的软件栈如PyTorch等进行安全漏洞检测，根据规范按时修复漏洞。
- 模型态势感知与事件响应：记录模型相关日志，如访问日志，审计异常访问。监控模型行为，识别异常，如高频请求、恶意请求、异常输出等。针对不同的事件，进行调查与响应处理。

3.7.3 华为云大模型安全解决方案实践

华为云大模型安全解决方案包括五大关键模块：大模型环境安全、语料数据安全、模型内生安全、模型推理安全、大模型安全运营系统。

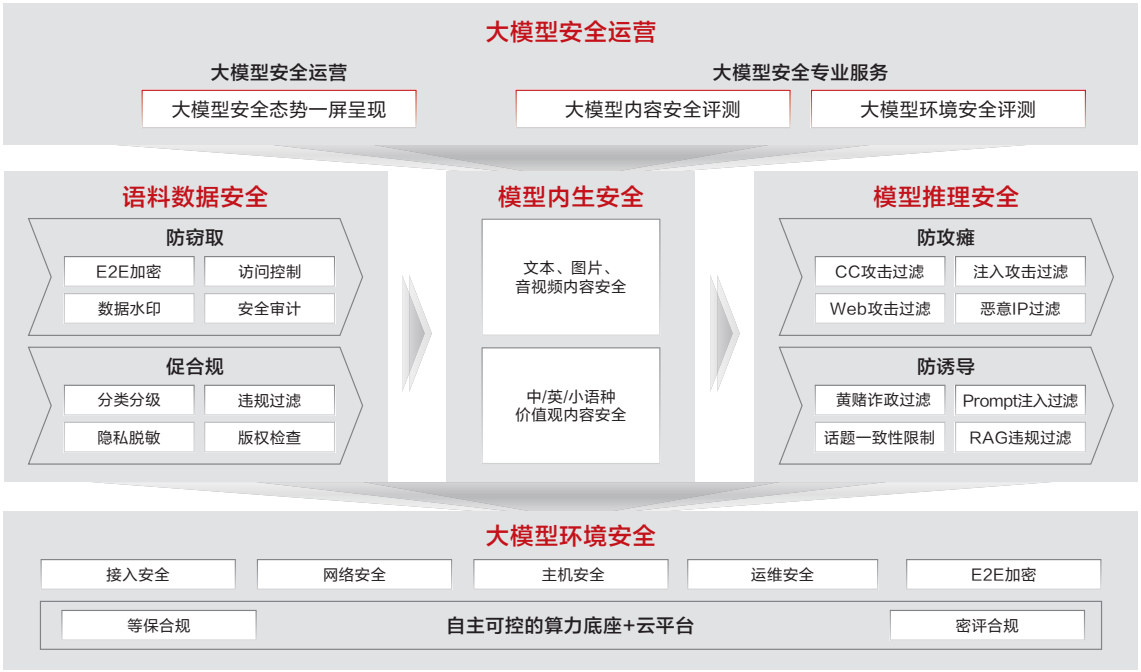


图122 华为云大模型安全解决方案实践

大模型环境安全：确保AI训练和运行环境的隔离与防护，防止外部威胁侵入，重点是构建覆盖网络+负载+应用+账号环境的威胁检测和防御能力，并对数据加密状态做到E2E可视化管理。包括但不限于网络隔离、访问控制、安全审计等措施，构建一个安全可控的AI训练和运行环境。

语料数据安全：通过数据加密、访问控制、数据脱敏等措施，保护训练数据的机密性和完整性。同时，确保大模型训练过程的版权保护和隐私保护。

模型内生安全：聚焦于模型本身的鲁棒性和抗攻击能力，减少模型被操纵的风险。这包括模型的自身对抗性训练等技术，提高模型在面对恶意攻击时的防御能力。

模型推理安全：在模型部署阶段，实施动态监控和防护，确保推理过程的安全。这包括提示词注入攻击、话题一致性限制、违规内容过滤等。同时，通过访问控制和权限管理，确保只有授权的人员和系统能够使用模型进行推理。

大模型安全评测：AIGC类大模型业务需要满足《生成式人工智能服务管理暂行办法》、《互联网信息服务算法推荐管理规定》、《互联网信息服务深度合成管理规定》的相关要求，对生成式人工智能进行内容评测、安全性评测并按要求进行合规备案。

04

AI-Native 技术在各行业领域的应用

4.1 AI-Native行业垂直应用

4.1.1 金山办公实践案例

1) AI 技术在金山办公的发展

随着AI技术的飞速发展，办公行业正经历着一场前所未有的革命性变革。金山办公，正积极利用AI技术重新定义传统办公的“写/存/管/用”四大核心场景。凭借三十多年的技术积累以及对用户需求的深入洞察，金山办公在2023年正式提出“AI x办公”战略，并发布WPS AI1.0版本，全面整合AI技术以提升办公效率，这一战略的提出，标志着金山办公全面拥抱AI技术，为用户带来更加智能、高效的工作体验。

在“AI x办公”战略的指导下，金山办公不断深化AI技术的应用，于2024年推出了WPS AI 2.0版本。这一版本的发布，进一步强化了金山办公在企业办公领域的布局，不仅提升了文档处理的智能化水平，还增强了数据安全和隐私保护，确保企业用户在享受AI带来的便利的同时，也能保护好自身的数字资产。

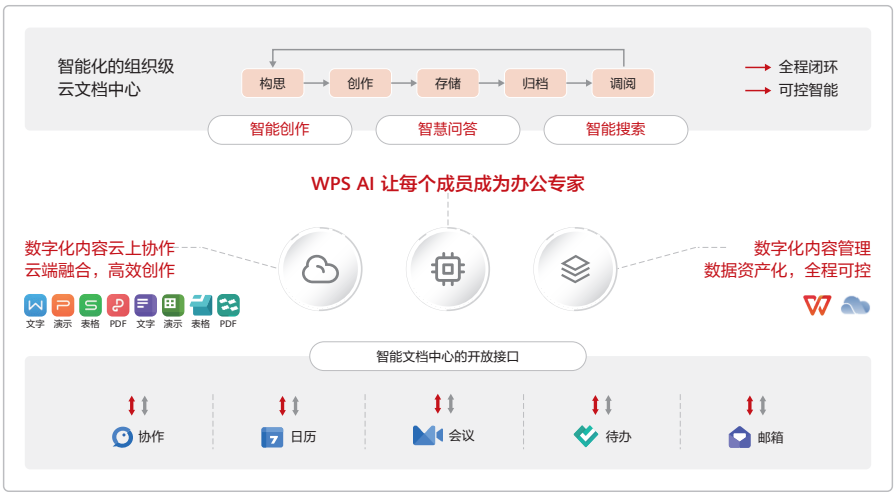


图123 WPS AI Native应用架构

未来，金山办公将持续用AI重构办公软件以及用户体验，持续优化WPS AI，帮助企业打造“企业大脑”，从而更高效地挖掘私有数据价值，提升生产力。

2) AI Native案例 - 企业AI知识库的实践

长期以来，金山办公一直致力于解决企业私域数据的有效管理和再利用问题。随着企业数字化进程的加速，企业数据呈指数级增长，传统的知识管理方式已无法满足高效决策和业务创新WPS AI全景图的需求。因此，企业迫切需要新的知识管理策略来应对挑战。AI技术的出现，为构建能够自动处理和分析大量数据的智能知识管理系统提供了可能。

金山办公AI知识库，作为WPS AI的核心基础设施，通过集中管理企业内外的结构化与非结构化信息，打破信息孤岛。同时，借助AI技术（如NLP、向量化、知识图谱等），实现企业私域知识的智能问答与创作、智能搜索、自动推荐、快速阅读等功能，从而大幅提升知识管理效率。

i. 企业私域数据的特点

金山办公结合AI能力针对企业数据的几个主要特征海量文档、领域知识、数据安全、文档质量构建了全方位的新一代智能文档库产品:

图124 WPS AI DOCS产品架构

在众多涵盖的场景中，比较受客户认可的亮点功能如下:

- 智能问答: 基于自研文档解析和RAG技术实现的精细化问答，支持深度思考、图文并茂、精准溯源等特性，在企业语料的回答准确率均值达90% 以上。

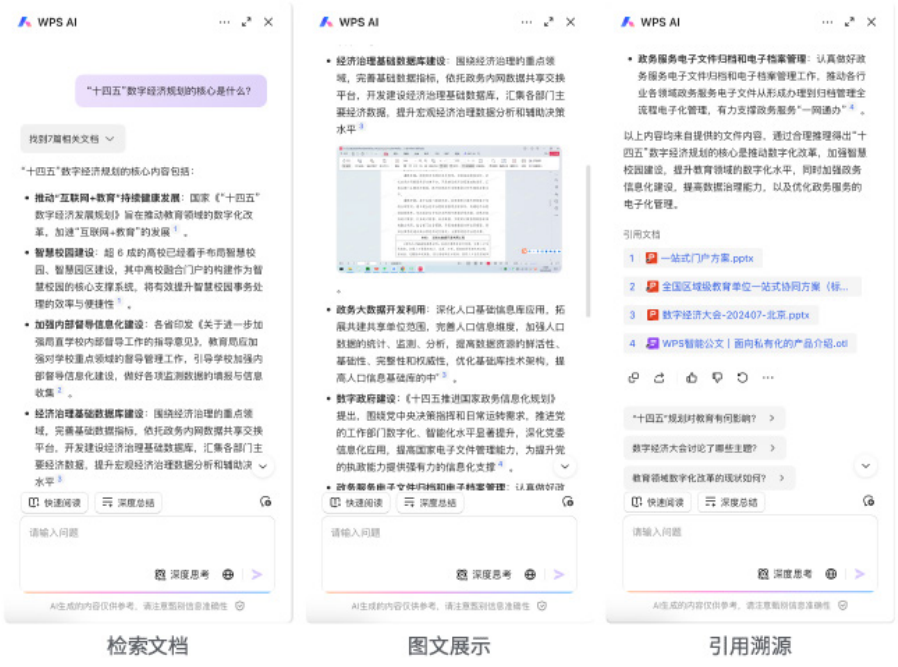


图125 WPS 智能问答应用

- 智能创作: 可基于企业内部模板作为生成大纲的参考，支持素材内的图片/表格，可根据创作内容生成脑图、流程图等配图。
- 知识图谱: 利用实体抽取和GraphRAG技术将文档转化为可视化的知识，提供知识查找、检索以及阅读等能力。

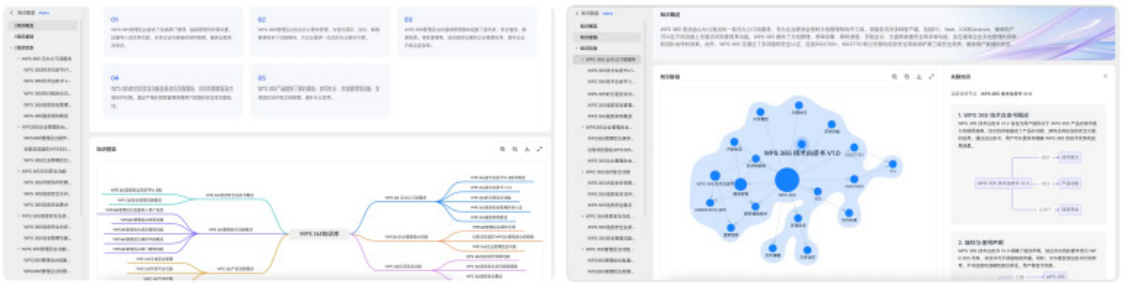


图126 WPS 知识图谱应用

- 智能抽取: 支持自定义字段按需提取，开放接入API并以json格式返回抽取结果。

ii. 企业AI知识库的技术挑战

- 亿级文档入库: 针对企业百万级文档规模，突破传统混合存储模式，采用「多引擎、存算分离」云架构。通过深度优化向量数据库，在亿级文档中实现400ms平均查询响应和98%以上召回率，为业务提供高并发场景下的可靠数据支持。
- 复杂文档解析: 针对企业多样的数据布局和文档结构（如嵌套表、流程图等），开发了KDC标准解析协议，为下游任务提供统一、精准的文档结构解析。
- 行业“黑话”: 针对企业内部特有的行业术语（“黑话”），开发了可扩展的问答系统，通过专业术语识别和提取技术，确保内容检索的准确性。
- 知识冲突: 针对企业文档不一致性导致的内容冲突。基于知识图谱技术开发了一套知识冲突检测系统，可以有效发现文档中的属性、数值、时间等冲突内容。

iii. 解决方案

通过「分层 x 开放 x 多模型适配」的架构设计来解决AI知识库在企业落地中遇到的挑战。

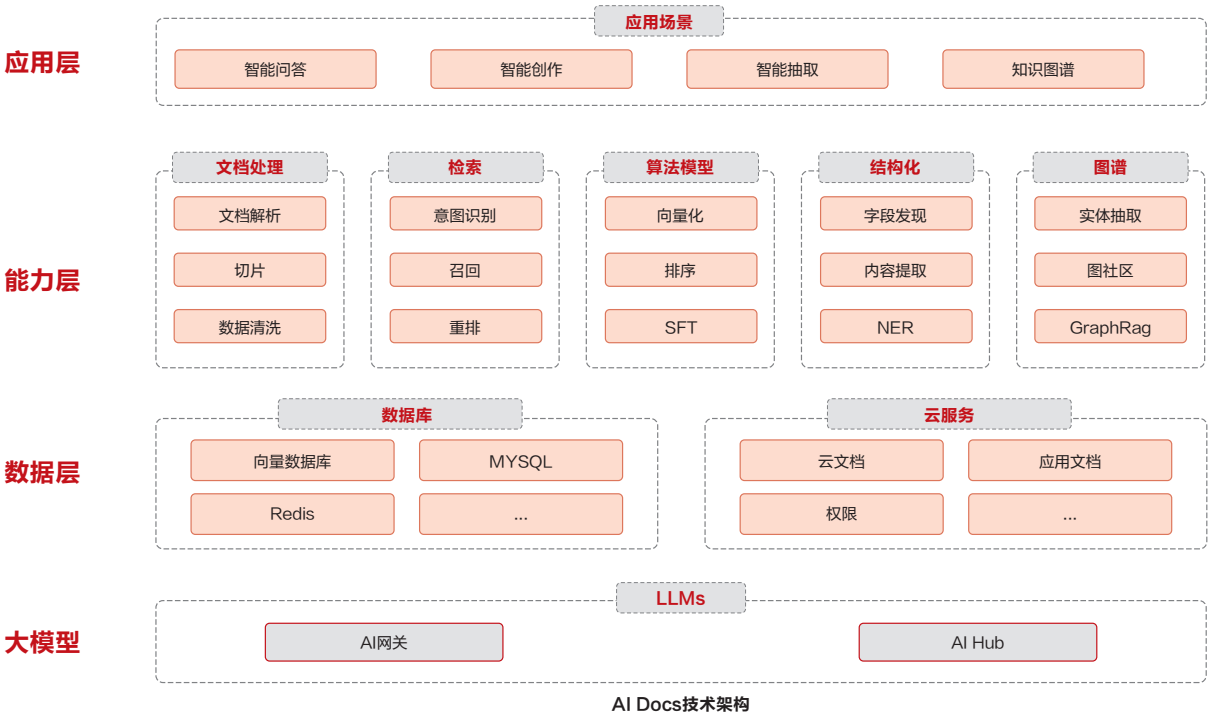


图127 WPS AI DOC技术架构

该架构具备以下特性:

- 精准解析: 集成版式还原模型, 文档解析准确率达95%以上
- 安全可控: 基于WPS 365云文档体系, 提供99.99%稳定性以及行业级的权限管理体系
- 灵活适配: 支持公私网部署, 根据企业需求可支持切换大模型
- 深度AI化: 支持亿级文档的AI Native处理能力, 云文档创建即可参与大模型处理
- 开放赋能: 提供丰富的API和扩展能力, 轻松接入企业的业务系统

4.1.2 美宜佳实践案例

自1997年成立以来，美宜佳坚持以特许加盟为主要发展模式，专注便利店业态。截至2025年3月，美宜佳全国门店数量已超38000家，覆盖全国22个省市，230多座城市，每月服务超过2.5亿+人次顾客。在生根并深耕便利店行业的28年里，美宜佳积极拥抱时代变化，主动探索数字化转型，发掘新的增长点，一步步构建起属于自己的门店网络，也逐步积累下广泛的消费者基础。

零售业是一个充满机遇和生命力的行业。当前，零售行业正处于一场由科技驱动的变革浪潮之中。一方面，物联网、人工智能等数字技术发展迅猛，不断渗透到零售业的各个环节，推动线上线下消费模式深度融合。另一方面，消费者对个性化体验和服务质量的追求，也促使零售业进行自我革新，以适应新时代的消费需求。正是在这样的市场背景下，数字化门店的概念应运而生。美宜佳凭借对市场风向敏锐的洞察力，在发展过程中迅速反应，致力于通过数字化升级与创新，提升终端竞争力，打造美宜佳数字化零售生态。

1) 数字化门店打造，推动业务升级

依托数字人技术、AI大模型与鸿蒙生态，美宜佳在购物体验、商品运营、门店经营及全域营销等多个维度上实现了突破与创新，为消费者打造了集互动、便捷与个性化于一体的购物空间。

图128 智慧门店总体设计蓝图

i. 6.0新形象门店升级

2024年7月，美宜佳正式对外发布了6.0新形象门店，进一步追求在数字化基础上的“增收、降本、协同”3大目标，让终端经营更便利。助力能耗降低20%以上，成本减少8%左右。门店能够在一定程度上减少开支，达成节能与降本的双重成效。

基于统一的物联网架构，可以快速支撑十万店与配套产业设备快速接入与稳定运行，通过智能控制实现节能与增效，支撑后续多业务场景的持续创新，实现“让供应商跟着美宜佳标准供货，而非美宜佳适配供应商标准”。

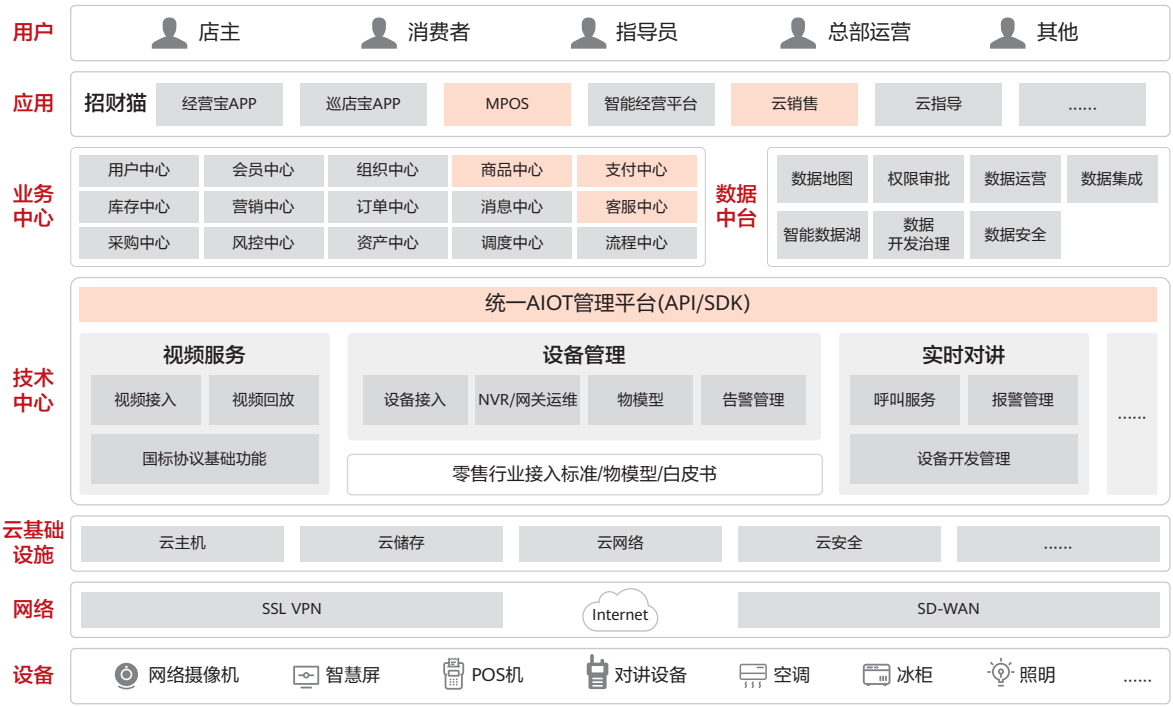


图129 智慧门店物联网架构概要方案

ii. 数字店员引入

美宜佳创新性地引入数字人技术，为顾客打造出能够提供“专属顾问”服务的数字店员。该数字店员能够依据购买历史和偏好，为顾客提供精准的推荐和个性化建议，使购物过程变得愉快。以大模型技术为基础，数字店员可以实现24小时、多语言、富情感、不间断服务，有效保持运营连贯性，将有力提高销售效率。这种人工+智能模式不仅能降低运营成本，还将显著增强顾客体验，打造更有“温度”的便利店。

华为云MetaStudio提供语音转文本、将大模型生成的文本答案，实时生成语音和数字店员视频，传输到终端进行播放等服务。

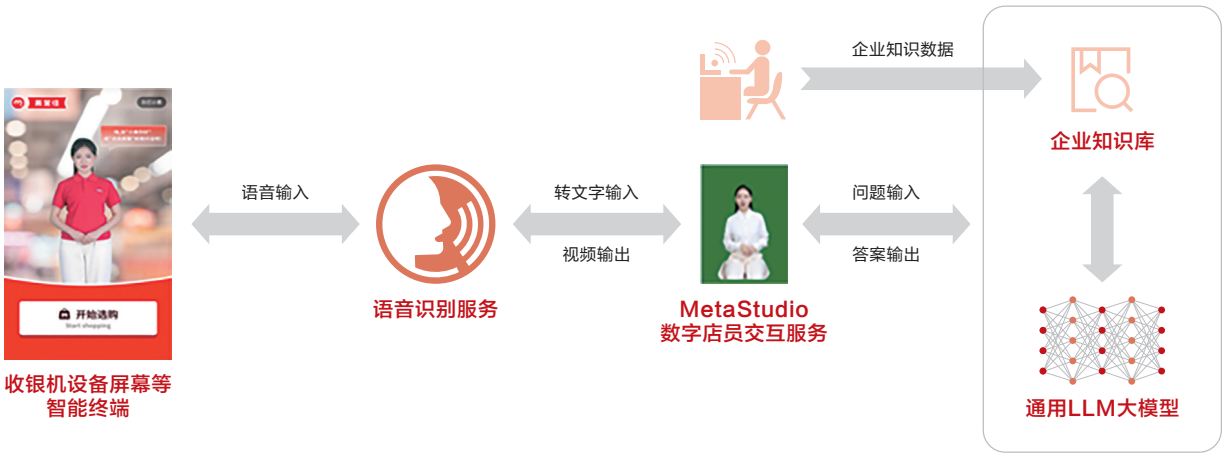


图130 数字店员产品交互逻辑

iii. 智能商品运营

在商品运营方面，美宜佳凭借大数据与AI算法，使全面智能化得以实现。通过对商圈数据和消费信息的精准分析，灵活自动地调整价格，在维护品牌形象的基础上，实现门店收益的最大化；借助AI算法连接供应链，在门店的进销存环节通过任务与物联网数字化货架更好地管理库存与门店陈列，使商品充足饱满，进而提升消费体验。

图131 智能陈列产品场景

iv. 打造全域营销

在全域营销领域，美宜佳把门店构建成了线上线下无缝衔接、融合共生的多维私域运营场所。借助线上直播、信息流广告等新颖手段，助力消费者更精准地匹配自身需求。与此同时，充分运用智慧终端屏的互动广告以及个性化推荐系统，依照会员的购物习惯和偏好为其推送专属优惠信息与产品推荐，提高用户的忠诚度和复购率。借力合作伙伴的全域资源，美宜佳达成了全方位、多触点的曝光，进一步拓展了品牌影响力，助力品牌市场份额的不断增长。

图132 智能营销产品场景

4.1.3 值得买科技实践案例

1) AI Native的购物助手实践

目前，消费信息的过剩已经成为一个事实。很多场景下，用户会需要消耗大量的时间和精力进行消费决策，包括信息的收集、整理、对比和甄别。AIGC技术某种意义上加剧了信息的泛滥。这对用户的消费体验产生了很大的影响。

同时，随着AI技术的发展，使得AI可以更好地理解消费世界，从而更好地帮助用户去理解消费内容中，其他用户的真实体感和商品特性，从而为用户提供更好的消费决策辅助信息。值得买科技基于AI能力，构建了“小值”购物助手。

2) 基于AI的购物助手架构

每一次新的技术革命，都带了一种全新的“力量”的供给。蒸汽机带来了全新的“物理力量”的供给，互联网带来了全新的“连接能力”的供给，而大模型的发展带来了全新的“智力”的供给。这种全新的能力供给，需要和行业相结合，从而发挥更大的力量。

消费领域的AI应用存在两个基本的要求：

- 实时性要求：消费信息应用的实时性要求很高，每天都有大量的消费经验和体验信息产生，同时商品的促销、价格等信息波动频率高（大促期每小时变动）；这些信息都需要能够快速地应用于对用户的服务。
- 事实性约束：消费信息里包括了大量的“事实”信息，例如商品本身的规格、参数、描述等等。消费领域的应用需要保证基础事实的准确。

而这些是无法通过单纯的大模型能力来解决的：

- 大模型的训练存在周期，通过实时数据进行持续频繁训练，通过“知识压缩”的方式来学习实时信息，其成本和时效性都无法保证应用的需求。
- 大模型本身的幻觉问题，是其数学机理所固有的，无法彻底避免和解决；

因此，必须在大模型能力的基础上，构建完整的应用框架，通过检索增强、思维链等方式，来完成应用的建设：

- 能够更高效地将实时消费信息体现在用户侧；同时，由于目前消费信息本身存在噪声，且很大一部分的内容是通过类似视频、图片的方式来表达的；而大语言模型和多模态模型，可以有效地理解用户的表述，理解多模态内容中的信息，使得值得买可以利用AI能力对消费信息进行预处理，提供更加准确的信息的同时，还可以避免对信息的重复计算，从而提升效率和降低成本；
- 通过大语言模型的思维链技术，能够更好地理解用户的真实需求。大部分情况下，用户表述的需求是一种场景化的描述，这就需要通过大模型对用户的需求进行分解，并且调度不同的工具来进行任务的执行。对于执行的结果，通过大模型来判断结果是否满足用户的需求，并且进行进一步的处理。

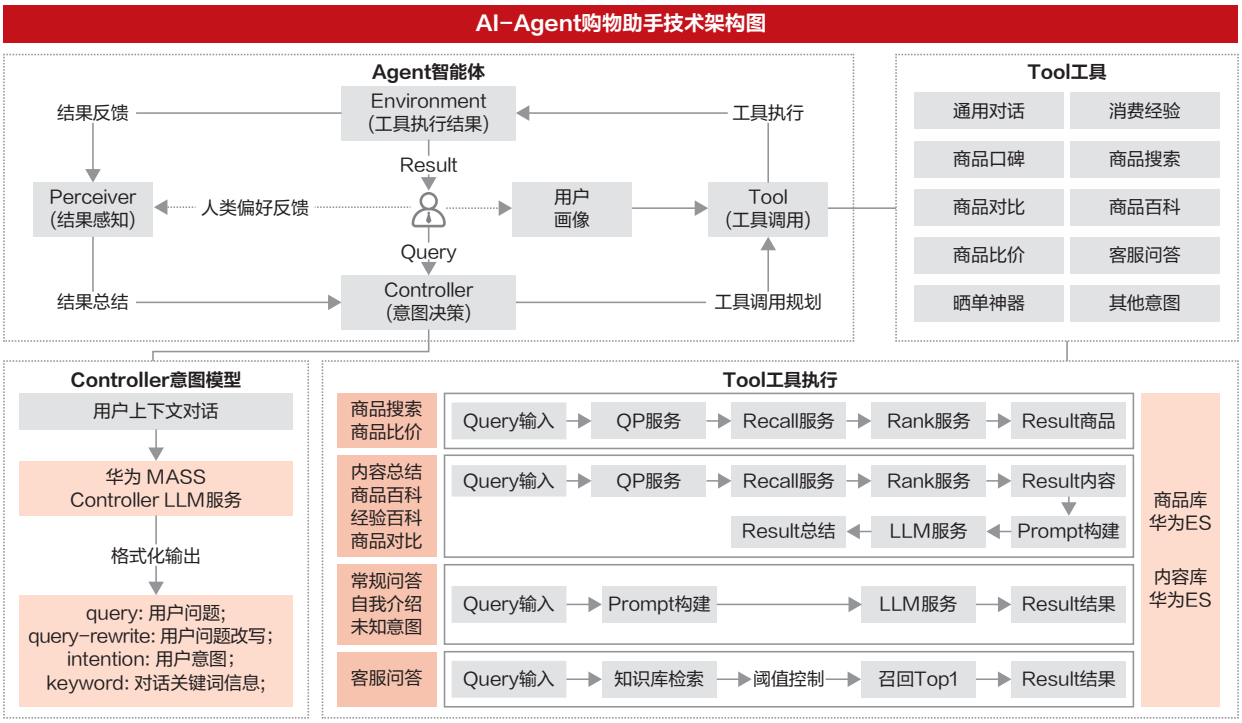


图133 值得买购物助手技术架构

基于这样的分析会发现，“小值”购物助手的整个体系，都需要基于AI来进行构建：

- “小值”购物助手的整个用户流程，都通过AI进行控制。AI会理解用户的需求，并且分解任务，选择执行任务的工具，构建工具参数，并且实际调用工具来解决问题。对于工具产生的结果，通过AI来判断结果是否满足用户的需求，从而决定是否返回结果，或者进一步使用其它工具来修正结果。值得买科技基于自有的消费数据对AI基座进行微调，形成自己的主控大模型，保证多工具分解的准确性和结果的稳定性。
- 将行业know how和AI能力相结合。值得买科技将用户的“消费行程”分为Touch、Seek、Taste和Enjoy四个阶段。并且针对四个阶段设计了通用对话、消费经验、商品口碑、商品搜索、商品对比、商品百科、商品比价、客服问答、晒单神器等9个专用工具。这9个专用工具通过“信息检索”和“大模型”的结合来完成专项任务。
- 消费内容数据库和商品数据库基于消费领域的特点（信息实时性和事实性要求高），海量的消费数据是模型训练、检索增强的基础。针对全域消费内容的筛选和处理，成为用户体验的关键点。值得买科技根据消费领域特点，基于AI能力构建了AIUC引擎，对全域消费数据进行阅读和筛选，从而形成实时的消费内容库和商品库，支撑整个购物助手的对外服务。

基于这样的设计，“小值”购物助手能够深度理解用户的意图，分解任务，执行任务，并且通过任务结果的判断来进行进一步的修正。

值得买科技认为，AI-Native并非仅仅是基于AI能力来打造产品，同时也必须拥抱AI所驱动的新的生态体系。随着AI能力的发展，人们的信息获取习惯已经开始发生改变，大模型产品已经成为很多人的信息获取入口，正在向决策入口的角色转变。而mcp协议的出现，使得AI和物理世界之间可以更加顺畅地进行连接；A2A协议的出现，使得agent之间的协作变得可能。AI生态中，每个角色应该以开放的心态，拥抱新的生态结构。

也基于此，值得买科技从构建“小值”购物助手伊始，就开放了自己在消费领域的各项能力，帮助更多的AI生态伙伴，更好地服务自己用户的消费需求。这也使得值得买科技可以更全面地接触到更多的用户，理解自己平台之外的更多消费者的需求。

AI Native，不仅仅是单个产品的构建，还是整个AI生态的构建。

4.1.4 汉得信息实践案例

AI 赋能制造企业产销S&OP决策会议

1) 业务痛点

供应链计划平台上线前，已有的信息系统仅仅是对计划结果进行简单线上管理的工具，月周日计划均由人工线下排出，没有进行有限资源模拟、自动化程度低，计划可执行性不高。主要问题包括：

- 计划体系业务流程线下环节多，存在管控性漏洞，缺少产业链计划协同联动能力，上下游计划脱节；
- 产销协同缺少数据应用的支撑，存在信息不同步、部门协同难、资源靠博弈、决策效率低、执行易脱节等问题；

2) 解决方案

核心数据架构：

图134 汉得信息S&OP全流程智能决策体系参考架构

i. 构建了月度需求计划-产销协同计划（S&OP）-周度需求计划-日滚动计划多层级科学计划体系，线上拉通从需求接收到计划下达的核心业务流程，提升生产组织管控精细度与敏捷度；

ii. 产销协同层面，基于大语言模型与RAG应用，构建S&OP全流程智能决策体系，从“经验驱动”升级为“数据-知识双轮驱动”，主要实现内容包括：

- 会前资源准备：AI+运筹优化资源平衡，RAG技术构建知识经验库，实现需求与供应智能匹配；
- 会中管理决策：模块化会议议程，多维产销存分析，实时What-if模拟，智能异常识别，自动生成会议纪要；
- 会后跟踪闭环：智能任务分派与追踪，超时自动提醒，确保100% 任务执行履约；

3) AI 应用场景和收益

基于大语言模型与RAG应用，构建S&OP全流程智能决策体系，从经验驱动升级为数据知识双轮驱动。

- 敏捷智能决策：建设AI 助力的产销协同S&OP决策平台，实现跨部门数据透明与协同，从战略决策到执行形成闭环，帮助制造企业构建全局视角，提升S&OP效率与准确性，真正实现了从“能用”到“智能”的飞跃；
- 整体计划流程优化：优化计划业务流程，横向贯通了需求计划、主机计划、结构件/产业链计划、采购/下料件计划，集成CRM、APS、ERP、MOM等系统纵向拉通了年/月/周/日计划；

4.2 华为云AI-Native行业应用实践

大模型技术在各个行业领域的应用正逐步深入，从医疗、金融、教育，到交通和零售，大模型以其独特的优势，正在颠覆传统服务模式，显著提升行业运营效率，并为日常生活带来前所未有的便捷。无论是在疾病诊断、金融风险管理，还是在个性化教育、智能交通控制，乃至销售预测和自动驾驶，大模型的身影无处不在。它们正成为推动这些行业不断前行的强大引擎，不仅优化了服务流程，还为各领域注入了新的活力和创新动力。

4.2.1 医学大模型行业应用实践

在医疗行业数字化转型的浪潮中，医疗行业客户可依托华为云盘古大模型的先进技术底座，通过“知识+数据”双轮驱动，构建覆盖医疗全场景的智能化解决方案。面对检验数据规模激增与临床决策效率提升的双重挑战，医疗行业客户能够以盘古大模型L0为基础，通过融合超过1TB的医疗行业专有数据及场景标注数据，来打造深度适配医疗场景的L1模型。盘古医学大模型不仅强化了医学知识推理、医疗信息结构化处理等核心能力，更可通过与华为云ASR语音识别、TTS语音合成、OCR等前沿技术融合，来构建覆盖“数据治理-智能分析-临床决策”的全链条智慧医疗平台。

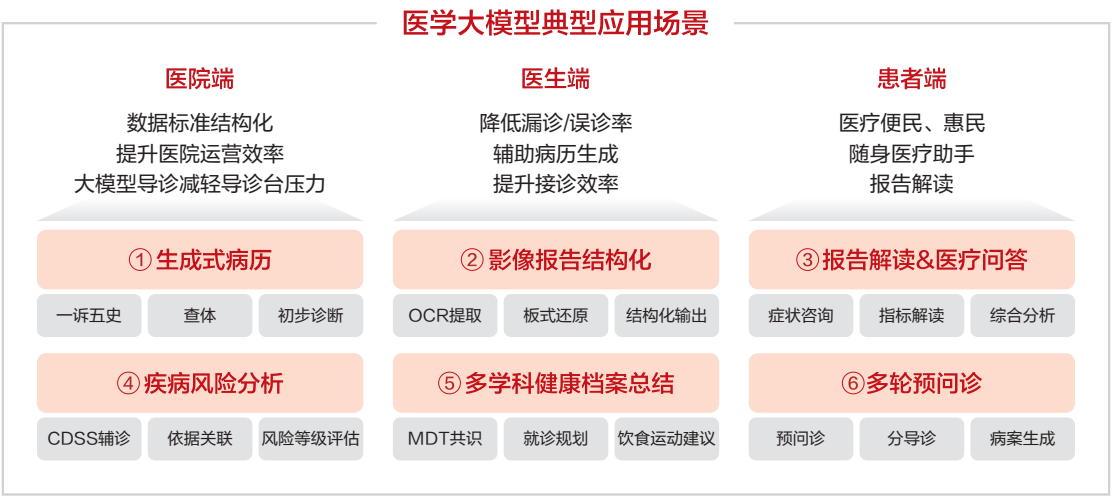


图135 盘古医学大模型

在场景化应用布局上，行业客户往往会聚焦“医院-医生-患者”三大终端来打造覆盖全医疗流程的解决方案矩阵。基于盘古医学大模型，可打造如下重点场景：面向医院端，通过检验数据智能治理系统实现多源异构数据的标准化整合，建立全院级医疗数据资产中心；针对医生工作场景，开发AI辅助诊断系统自动生成结构化报告，并结合患者历史数据提供个性化诊疗建议，使医生日均处理报告量提升3倍；在患者服务层面，构建智能健康档案系统，基于检验指标自动生成健康评估，并通过多轮对话引擎实现检验结果的通俗化解读。

面向未来，医学大模型会持续深化在医疗垂直领域的场景创新。配合业内头部客户通过构建“医疗大脑”知识中枢，打通检验数据与诊疗方案的知识闭环；加速行业客户开发自适应学习系统，实现模型在区域性常见病、罕见病领域的自主进化能力；最终形成覆盖预防、诊断、治疗、康复全周期的智慧医疗生态，推动精准医疗向个性化、前瞻性服务阶段跨越发展。

1) 大模型在医学行业的技术挑战

- 数据孤岛与碎片化：当前医疗行业面临严峻的数据治理挑战，各大医院内部运行的HIS、LIS、PACS等系统形成了严重的数据孤岛，超过60%的非结构化文本数据（包括病历、检查单等）分散存储且缺乏标准化处理。
- 临床实践面临双重效率瓶颈：医师日均处理文书时间达3.2小时（中国医师协会数据），门诊误诊率维持在8.4%高位。更严峻的是，现代诊疗中单患者检测指标已超50项，涉及血检、影像、基因等多模态数据，传统方法难以实现跨专科知识整合。
- 医患信息不对称问题突出：患者普遍存在专业术语理解障碍，而门诊沟通时间中位数不足8分钟。在慢性病管理方面，随访率仅37%，院外健康数据断层严重。

盘古医学大模型通过三大技术创新来攻克这些难题：首先，构建医疗专用的Tokenizer，专业术语识别准确率达98%；其次，通过高质量数据整合及配比训练，显著提升长文本病历的分析能力；最后，创新“知识蒸馏+强化学习”训练框架，使模型在疾病分析、报告解读、病历文案生成等多个场景达到90%以上准确率。

在医疗服务模式转型方面，医学大模型在过去的几年中已达成多个最佳实践，以某医疗行业ISV客户为例，其通过医学大模型搭建的应用在某试点医院构建起“诊前-诊中-诊后”数字化闭环，将患者满意度从65%提升至89%，同时通过智能分诊系统使基层医疗机构接诊能力提升30%。

2) 医学大模型案例 – 医疗场景应用落地实践

基于医学大模型，行业客户可将医院检验、检查等多元医疗数据进行系统化整合与标准化处理，构建完整的结构化数据体系。以医疗数据为核心要素，可进一步为医疗工作者提供精确、快速、便利的临床决策支持；为就诊者打造智能化、定制化、精准化的一站式医疗服务；为医疗机构数字化转型提供创新路径，全面推动智慧医疗、智能服务和精细化管理的发展。

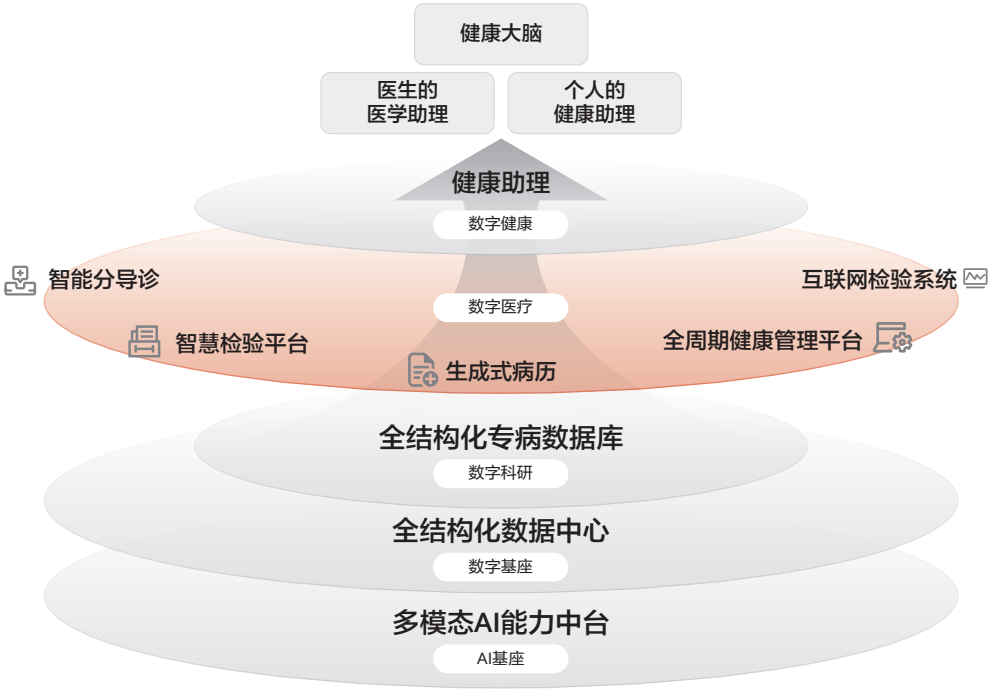


图136 基于盘古打造智慧医疗应用架构

基于医学大模型，业内客户已孵化多个场景应用，较为典型的场景能力如下：

- 医疗数据治理：基于全院结构化数据库，并依托医学大模型构建临床辅助诊疗系统，能显著提升医疗工作效率。盘古大模型能将院内碎片化的文本信息实现全结构化，加速院内数据治理。同时，盘古大模型的深度思考能力能进一步对患者特定时段内的全部检测指标进行智能分析，并重点标注关键指标和风险因素。有效帮助临床医师全面掌握患者健康状况，显著降低误诊漏诊概率。
- 智能病历生成：运用医学大模型的医学推理及医学文本凝练能力，整合患者当前症状与历史诊疗数据（包括病史、检查结果、用药记录、手术信息等），自动生成规范化医疗文书。该场景应用可大幅减少了医师的文书工作量，提高病历质量与标准化程度。某落地项目实践表明，该系统平均每日为主治医师节约2小时工作时间，整体工作效率提升达20%。
- 智能报告分析与解读：基于结构化数据库和盘古医学大模型领域推理能力，可开发面向患者的智能报告解读系统和全周期健康管理平台。智能报告分析能有效缓解患者焦虑，促进医患信息对称，提升医疗服务智能化水平。某落地项目中，客户结合盘古搭建“全周期健康管理平台”，构建了贯穿诊疗全流程、线上线下相结合的服务体系，提供智能导诊、健康教育和院后随访等多元化服务。该平台实现了对患者健康全周期和疾病全过程的双重管理，推动医疗机构向以患者为中心的服务模式转型，助力智慧医院建设，最终实现“医患双赢”的医疗服务新生态。

4.2.2 金融大模型行业应用实践

人工智能已成为驱动数字金融创新发展的核心驱动力，正推动银行业从“数字时代”向“数智时代”加速跃迁。根据中国人民银行《金融科技（FinTech）发展规划（2022-2025年）》，构建智能高效的服务体系成为支撑数字化业务发展的关键路径。2024年政府工作报告首次将“人工智能+”纳入国家战略部署，标志着AI技术将开启全行业深度应用的新纪元。

作为数字化转型的标杆领域，金融业不仅拥有丰富的人工智能应用场景，更具备完善的技术实施基础。为落实国家创新发展战略，金融机构正加速构建“AI+金融”的生态体系。在传统AI技术迭代成熟与金融业务经验持续积累的双重驱动下，智能金融已进入规模化发展的快车道。值得关注的是，生成式AI大模型的突破性发展正在开辟新的赛道，通过与传统AI技术的深度融合，推动数字金融向智能化、场景化方向实现质的飞跃。这种技术融合不仅拓展了金融服务的边界，更孕育出全新的业务模式与价值增长点，为行业转型升级提供了历史性机遇。

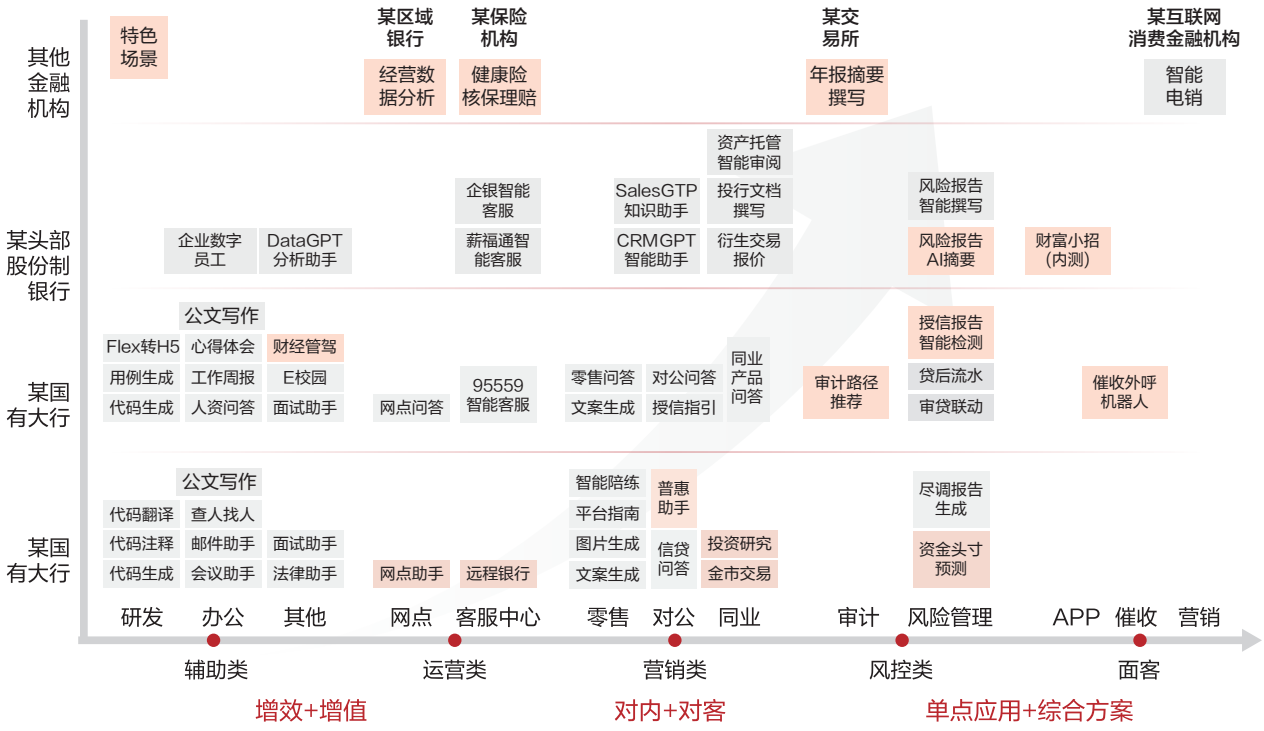


图137 金融行业大模型应用场景地图

华为云金融大模型基于“数智融合基座”，构建了覆盖金融全场景的智能解决方案体系。基础层整合五大核心能力：文档智能解析、商业智能分析（智能BI）、代码生成工具、数字人交互系统及知识图谱管理，实现金融数据到价值的高效转化。应用层聚焦八大核心场景，打造智能化金融对客服务、分钟级智慧信贷审批、全流程办公辅助及智能运营管理解决方案。

该模型深度融合传统AI技术与生成式大模型优势，在风险控制领域实现突破：通过多模态技术整合，将信贷审批时效从天数缩短至分钟级；智能风控矩阵覆盖贷前调查、贷中监控、贷后管理的全生命周期。目前已在金融文本生成、合规审查、量化策略等专业领域达到行业领先水平，为金融机构提供了从基础能力到垂直场景的完整AI赋能路径。

- 对客服务：大模型辅助内部客服人员通过生成推荐优秀话术提升服务质量，识别前期交流的意图及关键信息以供转接人工客服后参考，提示历史重复致电以辅助客服交流。在客服运营方面，模型提取客服工单要素，自动记录并辅助核验，识别转人工意图并统计以支撑智能客服语料优化，识别服务负面情绪和关键信息以支撑质检。在辅助营销中，模型基于客户交流的意图及关键内容识别客户潜在需求，支撑用户画像用于营销。直接对外服务方面，模型支持在线询价，询价后转到业务流程，同时提供闲聊服务。

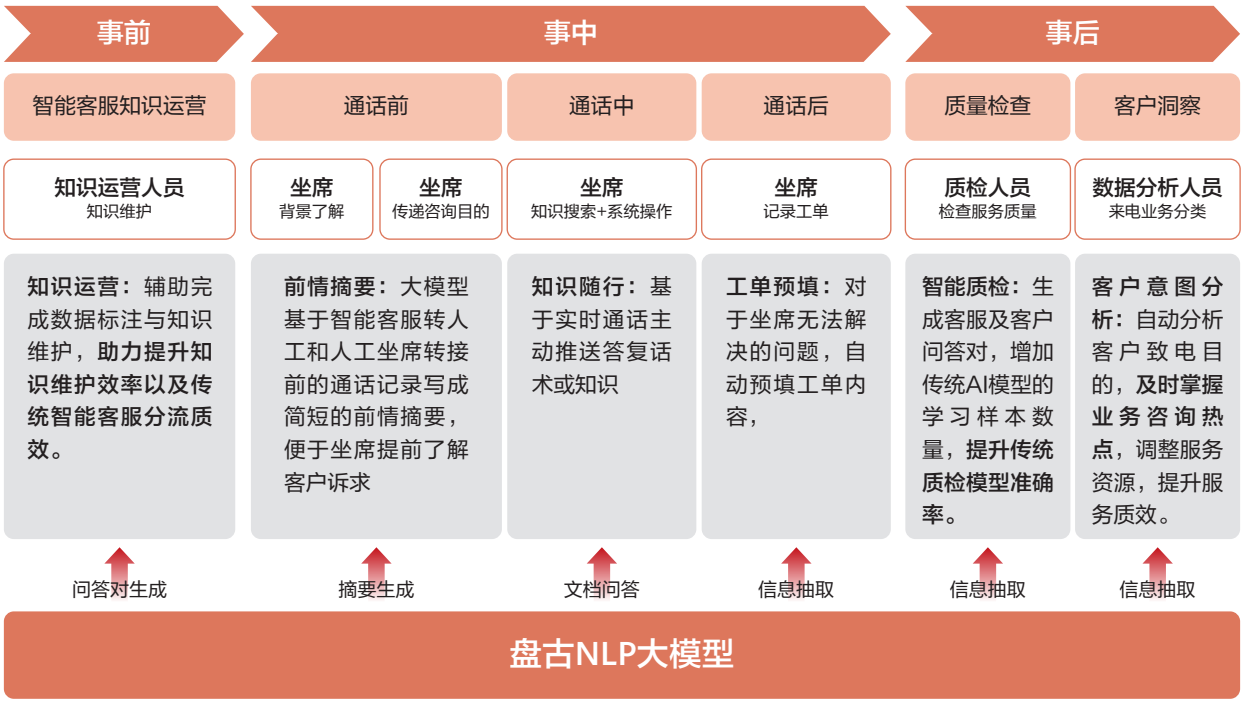


图138 大模型加持的智能客服业务流程图

- 智慧信贷：大模型提供信贷制度问答，迅速响应关于信贷政策的各类咨询，为贷款经理和客户提供详尽的制度解读，确保信贷活动的合规性。针对行内外流程和规范提供即时解答，助力员工快速掌握操作要领，提升工作效率。根据贷款申请的详细信息，生成格式规范、内容全面的审批文件，为信贷决策提供专业支持，缩短审批周期。

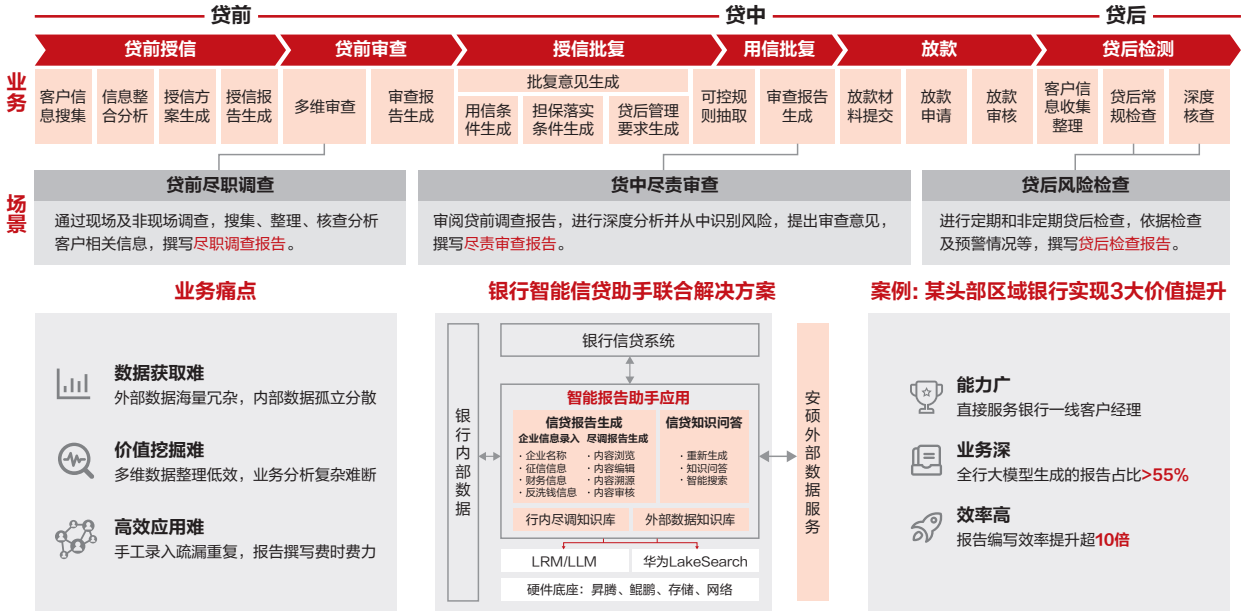


图139 大模型加持的智能信贷业务流程图

- 办公助手：大模型为跨国金融交流提供精准、即时的语言转换服务，确保信息在全球范围内的无缝对接。能够准确捕捉会议中的关键信息，自动生成详尽的会议纪要，为金融决策提供可靠的历史记录和参考。迅速解答各类金融问题，为员工提供即时的专业知识支持，增强团队的专业能力。即时响应员工对网点运营制度的咨询，确保工作流程的规范性和一致性。具备将自然语言查询转换为数据库查询语言的能力，使得非技术背景的金融专业人士也能轻松进行复杂的数据分析，从而洞察市场趋势，辅助决策制定。

图140 某头部大行办公助手方案

- 运营管理：大模型能够快速生成投研报告，从咨询、财报等资料中提取关键信息，为金融市场人员提供使用。内部网讯宣传稿生成方面，模型能够根据关键活动等新闻快速生成宣传稿，提升内部沟通和宣传效率。

4.2.3 气象大模型行业应用实践

气象预报是一个从数据收集到预报生成的复杂流程。首先，通过地面观测站、卫星和雷达等设备收集温度、湿度、风速等气象参数。随后，这些数据通过数据同化技术与数值预报模型结合，以得到符合实际大气状态的初始条件。利用高性能计算机，数值预报模型基于物理方程模拟大气行为，生成预报数据。这些原始预报结果经过后处理，包括偏差校正和统计分析，以提高预报的准确性。预报结果还需经过验证，与实际观测数据比较，确保预报质量。最终，预报产品以图形、文本等形式发布，服务于天气预报、气候研究和灾害预警等。传统数值天气预报模型虽然在气象预报中发挥了核心作用，但该技术已经达到了瓶颈，主要存在以下痛点：

- 数据利用效率低下：全球30颗气象卫星和200颗研究卫星每天产生20TB观测数据，但仅有5%用于模型计算。这源于数据格式不统一、存储传输能力不足，以及模型分辨率与观测设备不匹配等技术瓶颈。
- 算力需求爆炸式增长：传统数值预报单次运行需数千核时计算资源，欧洲某核心机构为提升精度，其模式分辨率提升导致计算规模呈指数增长。预计2025年将触及现有超算算力天花板，同时伴随数据同化和后处理等环节的连锁挑战。
- 时效性难以满足应急需求：完整预报流程需数小时，涉及数据同化、模型运算及结果验证等环节。这种延迟对暴雨、龙卷风等极端天气的实时预警构成制约，可能影响灾害应对黄金时间的把握。

针对气象行业中面临的问题，华为云推出了盘古气象大模型解决方案。

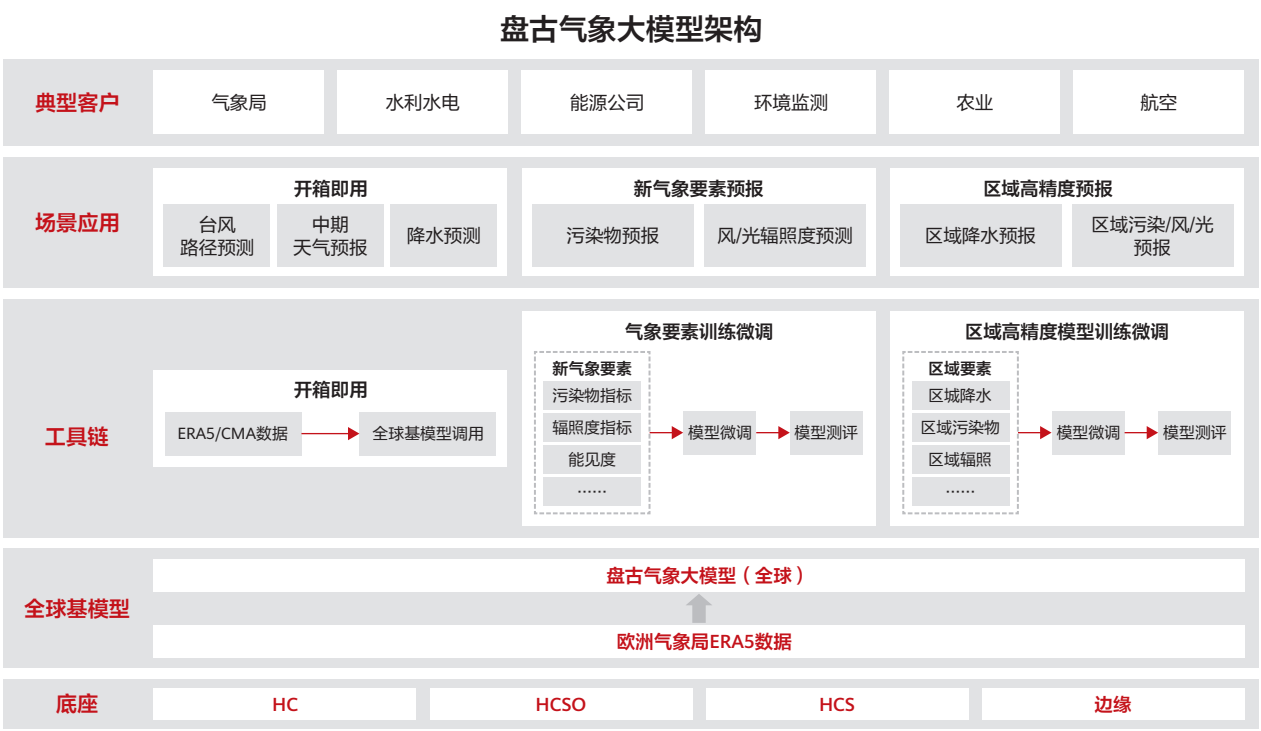


图141 气象大模型解决方案架构图

盘古大模型解决方案内置全球气象预报基模型，该模型采取43年的欧洲气象局ERA5数据训练。

1) 在工具链层面

- 开箱即用: 支持开箱即用的全球气象预报模型的训练推理能力; 支持中期天气要素预测、台风路径预测、降水预测场景
- 新要素模型训练微调: 可以支持可视化图形界面，低代码训练微调新气象要素大模型
- 区域高精度模型: 以全球气象预报模型作为边界条件，提供针对特定区域，提供高精度的区域模型训练和推理

2) 在场景层面

- 全球中期天气要素预测: 提供未来10天的全球中期天气要素预报，主要提供水平分辨率25km*25*km，1小时模型时间分辨率，垂直高度包括13个等压面（50-1000hpa）和地表，5种变量（重力位势、温度、风速、湿度、气压）。模型的RMSE、ACC等指标超过传统数值预报模型10%以上。
- 台风路径预测: 预测数天内台风移动的路径、强度变化和伴随大风等等状况。台风路径轨迹预测的准确率超过欧洲气象中心25%。
- 降水预测: 在中期天气预测的基础上，增加了对降雨量的预测，支持预报累计6小时降水。可结合集合预报共同使用预测累计6小时降水量的概率。降水预报Treat Score相比SOTA提示20%，可提前3天预报结果

大气污染物预测: 可预测全国6种污染物（PM2.5，PM10，O3，NO，SO2，CO），支持预测未来7天的AQI指标，精度提升10%。同时输出污染物传输影响值。

气象大模型通过“AI+气象”的融合创新，正推动行业服务模式发生根本转变。其核心价值不仅体现在突破性解决传统气象领域的数据利用率低、算力需求大、响应时效短等痛点，更在于构建起覆盖多领域的智能气象服务网络——通过算法驱动的精准预测能力，将气象要素与行业场景深度耦合，为防灾减灾、城市治理、能源转型等关键领域提供决策支撑，标志着气象服务从被动应对向主动预防、从通用预报向场景赋能的战略升级。

- 农业气象服务: 在现代农业生产中，利用气象大模型，农业管理部门能够获得更加精准的气象服务支持。通过高精度的天气预测，气象大模型可以帮助农民合理安排播种、施肥、灌溉和收获等农业活动，从而提高农业生产效率。例如，当预测到即将出现强降雨时，系统可以提前预警，提醒农民采取防护措施，减少灾害损失。此外，模型还能结合气象数据分析病虫害发生的可能性，帮助提前制定防治方案，实现智慧农业的目标。
- 新能源发电优化: 在新能源发电领域，气象大模型的应用能够显著提升风能、太阳能等可再生能源的利用效率。通过对风速、光照强度等关键气象因素的精确预测，模型可以帮助新能源企业优化发电计划，减少因天气变化导致的发电波动。例如，在风电场运营中，系统可以根据实时气象数据预测未来的风力变化，指导电力调度中心合理分配电力资源。此外，盘古气象大模型还能为极端天气提供预警，保障新能源设施的安全运行。
- 航空气象服务: 在航空运输中，气象条件直接影响飞行安全和效率。气象大模型通过高精度的气象预测，为航空公司和机场提供优化的航线规划和天气预警服务。例如，当预测到航线区域可能出现强对流天气时，系统可以提前建议调整航线，确保飞行安全和准点率。此外，模型还能为机场提供降水、风速等关键气象数据支持，优化机场运行管理，减少因恶劣天气导致的航班延误。
- 极端天气预警: 在防灾减灾中，气象大模型能够显著提升极端天气事件的预测能力。通过对台风、暴雨、洪水等灾害的精准预测，模型可以为政府和相关部门提供科学的决策支持。例如，在台风来临前，系统可以准确预测台风路径和影响范围，帮助政府提前部署救援力量，减少灾害损失。此外，模型还能为公众提供个性化的预警信息，提升社会整体的防灾意识和应对能力。

4.2.4 矿山大模型行业应用实践

全球矿业正经历数字化转型的关键时期。在能源结构转型与“双碳”目标的双重驱动下，矿业企业亟需通过智能化技术重构生产范式。矿山AI大模型的创新应用，标志着矿业从传统经验驱动向数据智能驱动的范式革命，为行业可持续发展开辟了新路径。

煤炭行业推进智能化建设，依赖人工智能技术的支持，但传统单场景小模型方案存在诸多问题，制约了矿山智能化、规模化建设的发展。以矿山智能应用的业务视角分析单场景小模型方案，存在以下问题：

- 模型泛化能力不足：传统单场景模型缺乏跨矿山移植性，不同矿区应用时精度显著下降，难以规模化复制。
- 动态工况适应性弱：模型在开采条件变化（如地质结构突变、设备工况波动）时，难以保持精度和扩展性。
- 数据安全隐患突出：线下开发模式导致矿企核心数据在传输、存储环节面临泄露风险，威胁生产安全。
- 开发效率低下：采用碎片化“作坊式”开发流程，缺乏知识复用机制，相同功能需重复开发，项目周期长达数月。
- 技术门槛过高：全流程依赖AI专家经验，开发者需同时掌握算法优化、行业知识、工程部署等多领域技能，人才缺口显著。

针对以上挑战，华为推出矿山大模型解决方案，采用“1+4+N”架构体系。其中“1”为统一AI基础平台，通过分层解耦的云原生架构实现跨矿区协同；“4”构建智能感知、数字孪生、决策优化、安全防护四大核心能力模块；“N”支持多场景智能应用落地。方案创新应用无监督学习方法，在保障数据隐私前提下实现跨矿区知识迁移，解决传统模型泛化性不足问题。通过预训练大模型与行业知识融合，可快速适配采煤、运输、安全监测等多样化业务场景，显著降低开发周期与技术门槛，为矿山智能化建设提供安全高效的解决方案。

图142 矿山大模型解决方案架构图

“1”为矿山一站式AI平台：基于华为云构建，提供全流程大模型训练/推理服务，支持算法管理、多框架开发、模型统一管控及弹性资源调度，实现AI开发全生命周期管理，显著提升模型开发部署效率。

“4”为核心能力层：

- L0层基础大模型：基于千亿参数规模预训练，集成视觉、预测、NLP、多模态四大通用能力，具备行业领先泛化性；
- L1层行业模型：融合煤炭行业知识（含百万级矿山图像数据），开发物体检测、语义分割等专业套件，支持授权调用；
- L2层场景模型：通过可视化工作流实现场景化模型训练，支持按需抽取模型结构、适配数据特征，产出可部署的推理模型。

“N”为场景化解决方案：通过标准化工作流机制，支持多业务场景模型开发。在数据安全框架下，用户可选择授权套件进行L2模型训练，实现从通用能力到垂直场景的精准适配，构建矿山智能化应用生态。

矿山大模型的优势在于它不仅能有效提升样本训练效率、降低样本标注的人力成本，还能与矿山业务应用深度融合，通过小样本快速训练出需要的场景化模型。同时，矿山大模型具有高泛化性和移植性，能适应矿山的不同业务场景。此外，矿山大模型实现了全栈自主创新，为煤炭行业智能化建设提供了综合解决方案。其构建的全场景数字孪生体系已赋能某集群矿山生产流程重构，通过多维度数据实时感知与智能决策，显著提升设备运维效率，实现高危作业场景人员安全管控，推动安全生产水平跨越式发展。

在资源开发领域，模型通过地质构造智能解析与开采方案动态优化，助力某集群矿区资源利用率大幅提升。其具备自主进化的安全预警系统，可精准识别多类风险场景，构建起分钟级灾害响应机制，树立行业安全管控新标杆。

目前矿山大模型已深度覆盖全国多个省级能源基地，适配井工矿、露天矿等全品类场景，形成单矿智能化改造成本优化效应。通过“模型即服务"的生态赋能模式，带动产业链上下游协同升级，打造可复制的矿山数智化转型方案，持续推动行业高质量发展。

4.2.5 政务大模型行业实践案例

目前我国的城市治理模式，需求侧主体主要有人民群众诉求和领导要求两方面。其一，人民群众的诉求，即民意。具有聚焦准、切口小、可参与度高，以及“短、平、快、新”的特点，全面、准确、及时了解社情民意。其二，城市管理者的要求。城市管理者在组织中扮演着至关重要的角色，具有决策、指挥和协调的作用，而群众的诉求是城市管理者要求的重大来源。

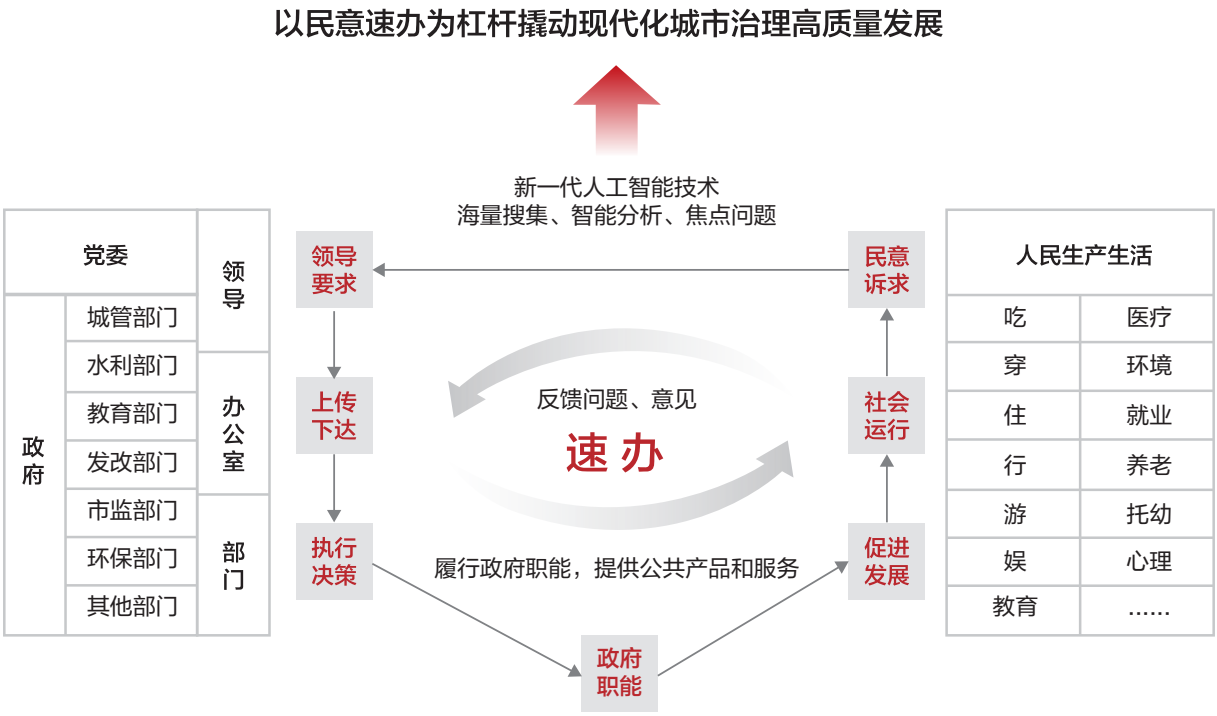


图143 人工智能赋能现代化城市治理方案

面对超大城市错综复杂的民意诉求，基层治理主体的传统治理手段已不能支撑诉求的高效处理，而以ChatGPT、DeepSeek等自然语言大模型为代表的新一代人工智能技术在民意诉求的搜集、分配环节均显现出强大的赋能效应，为后续民意诉求的处理、反馈提供了坚实的基础。

1) 政务大模型创新推动“民意速办”

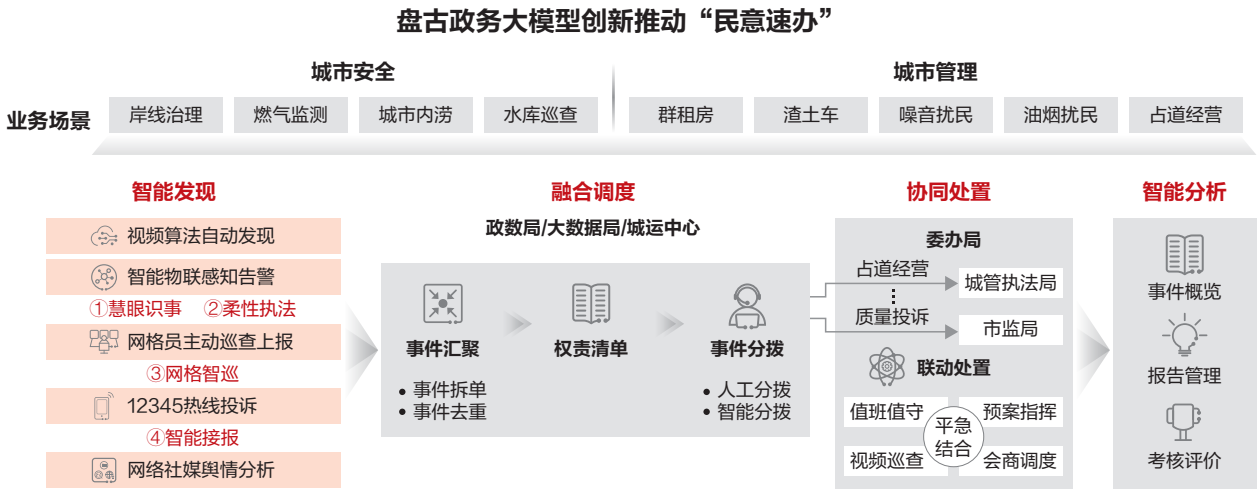


图144 大模型加持的民意速办业务流程图

i. 智能发现：智能接报 —— 让接报更高效，工单“一字不填”

12345热线整合多类事件渠道后，每日工单数量大幅增加，工单受理登记的质量以及热线接通率成为瓶颈，影响市民对城市的满意度。在工单受理中有超过10项高频字段需要填写，人工在长时间繁重的压力下很容易出错，同时，每个工单的处理仅填写工单内容就需要超过3分钟，这也制约了热线的接通率的提升，影响了市民的整体体验。

通过大模型技术可以实现热线智能接报，辅助人工自动填写各类事项的高频字段，正确率达到98%，实现工单受理“一字不填”，每单可节省近3分钟，能极大地提升热线接报质量和效率。

图145 盘古大模型政务工单自动填写样例图

ii. 融合调度：智能工单 —— 让工单快速分拨处置“一单到底”

在分拨阶段传统人工分拨依赖“老法师”经验，准确性和一致性无法保证，面临投诉工单被接收部门退回，甚至会出现多轮“踢皮球”的情况。

通过大模型智能分拨技术，将城市治理通用知识、本地特有知识以及实时动态知识整合成有机的整体，构建动态在线高效智能的知识处理系统，发挥大模型处理城市治理任务的独特优势，进一步提升分拨效率，同时将分拨准确率提升至90%以上。

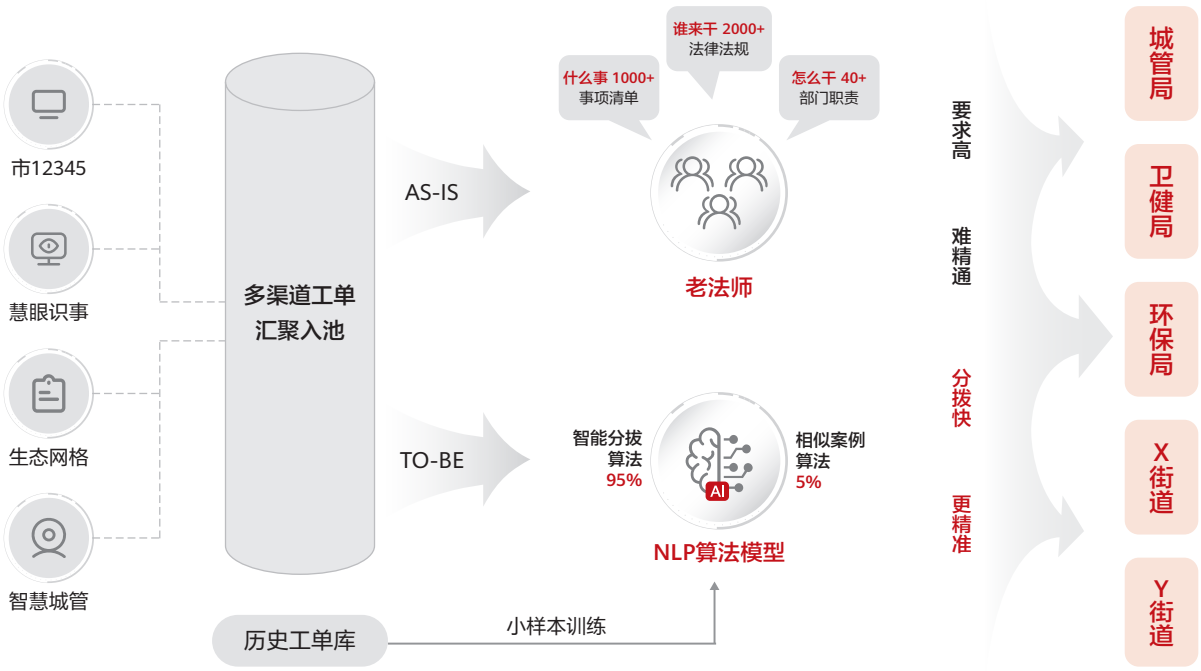


图146 盘古大模型实现工单智能分拨

iii. 协同处置：指挥中心 —— 让指挥更精准、更高效

民意速办分拨工单到处置单位，处置单位利用全域覆盖、资源丰富、高分辨率的指挥一张图上，汇聚救援队伍、物资储备库、避难场所等应急资源，结合实况地图等空间能力，实现图上部署和挂图作战。利用系统提供的结构化应急预案、任务调度等能力，使得指挥员能实时、直观地感受空间要素，实现智能化、精准化指挥。

同时利用融合通信等多种手段，解决现场信息采集和通信问题，实现现场指挥“看得见、听得到、连得通、指令下得去”，有助于提高城市治理事件实时调度处置能力。工单处置完成，把处置结果反馈给民意速办，形成闭环。

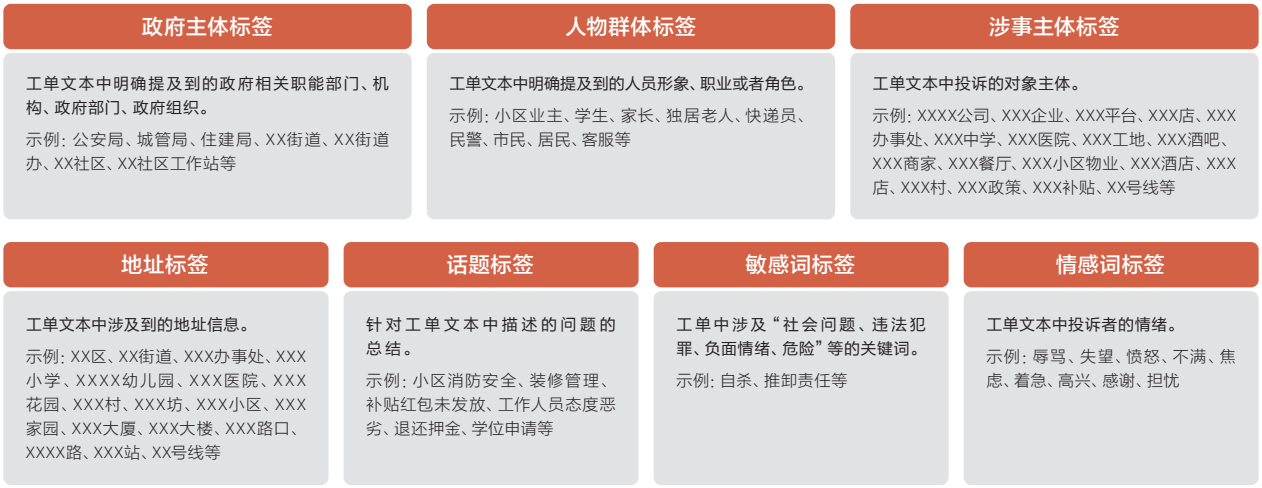


图147 政务数据标签体系

iv. 智能分析：智能标签 —— 让工单打标全面准确

数据分析首先需要形成可供分析的基础数据，民意速办工单的标签种类多，包括政府主体、人物群体、涉事主体、地址、话题、敏感词、情感词等七大类，同时又有数千类的事项类型，传统人工方式难以完成对每天大量的事件工单快速统一标注标签的工作。借助大模型的特征识别能力，可以快速对全部工单按照统一的标签体系完成全量标签的标注，形成高质量的基础工单分析数据，支撑对市民的民情民意分析、对部门的处置效能分析。

基于包含领域知识的工单效能分析模型、工单知识图谱，利用大模型的报告自动生成能力，可以快速生成高质量的民情民意报告，为城市治理提供高价值的决策依据。

2) 城市现代化治理实践

华南某区12345热线推出“民意速办”小程序平台，平台整合全市537个民生诉求渠道，在“民意速办”小程序平台上，市民只需要手机登录、输入诉求，通过智能派单应用，AI模型将根据投诉内容，进行承办单位判定，实现24小时自动无间断派单，让工作人员能及时跟进处理，智能派单准确率已达94%以上。

同时结合效能数字监管平台，通过流程效能监管、重点工单监管、效率监管、语音质检等模块，工作人员可以对工单的办理环节、办结分析、办理效率等多个维度进行全流程实时监管，及时捕捉热点、难点、堵点问题，实现热线效能监管。

通过汇聚的近百万条工单，挖掘4756个热点话题，分析诉求与诉求本身、责任主体、诉求主体、处置措施之间的内在联系，由此构建了知识图谱，从时间、空间、位置和人群四个维度动态精准感知热点，从安心、安乐、安康、安居四个指标构建居民福祉指数体系，并以此来助力城市治理，支撑科学决策。

05

AI-Native 技术的关键挑战

5.1 模型透明性与可解释性问题

随着AI模型尤其是深度学习模型的复杂性提升，如何理解模型的决策过程成为一大挑战。尽管AI能够提供高效的决策支持，但其“黑箱”特性仍然令许多行业和用户对其信任度产生疑虑。随着AI模型在各行各业的广泛应用，模型的透明性和可解释性问题也愈发显现。在许多关键领域，如金融、医疗、司法等，AI模型的决策往往需要被理解和解释，以便进行合规性审查或向用户提供足够的信任。然而，许多深度学习模型特别是生成式AI模型，往往被视为“黑箱”，即其内部决策过程难以理解或追溯。

解决这一问题的一个方法是引入可解释AI（XAI）技术。可解释AI不仅旨在提高模型的透明度，还要确保模型的决策过程能够被用户、开发者和监管机构理解。可解释性不仅有助于提高系统的信任度，还能够在模型出现偏差或错误时提供必要的诊断信息。此外，可解释性还可以增强模型在敏感领域中的应用，避免由于“黑箱”特性带来的法律和道德风险。

此外，类似DeepSeek等直接输出模型推理过程的做法也是增强可解释性的有效手段。例如，用户可以根据LLM大模型输出的推理过程数据回溯模型输出结果的合理性，对于输出错误的案例，可以分析发现模型推理过程中存在的薄弱点，进而有的放矢地对模型能力进行增强。

然而，要实现高度的可解释性，往往需要对各式各样的模型结构和数据特征进行深入分析，这对算法和计算能力提出了更高的要求。未来，随着AI技术的不断发展，如何平衡模型的准确性与可解释性将成为技术进步的一个重要方向。

5.2 模型安全治理挑战

随着AI在企业中的应用越来越广泛，模型安全问题成为了一个日益严重的挑战。AI模型在处理敏感数据时，容易受到外部攻击，如对抗攻击（Adversarial Attacks），这些攻击可能通过微小的扰动改变模型的预测结果，进而引发系统错误或安全隐患。与LLM应用相关的安全挑战包括模型供应链不可靠、智能体权限过大、提示注入攻击、无限资源消耗攻击、不当输出、敏感信息泄露等。因此，如何保障模型的安全性成为了AI原生技术发展过程中不可忽视的一环。

模型的安全治理要求从多个层面进行保障。首先，在模型训练阶段，应加强数据的安全性，确保数据来源可靠且符合隐私保护要求。其次，在模型应用阶段，应通过防御性算法设计，提升模型对抗攻击的能力，确保其在面对未知威胁时依然能够稳定运行。最后，还需要建立健全的模型监控机制，实时检测和修复潜在的安全漏洞，从而确保系统长期稳定与安全运行。

例如，华为云大模型安全解决方案通过多层次技术构建防护体系：在数据安全层面，采用用户自主掌控的密钥管理服务（KMS）对模型进行端到端加密，结合数据水印技术实现溯源追踪，依托机密计算环境保障数据处理过程可信；在开发安全维度，通过构建DevSecOps流程将安全防护前移至设计阶段，配合分级脱敏策略和全链路数据血缘追踪，实现从模型训练到推理的全过程审计与透明化管理；在运行防护方面，不仅通过自动化测试工具持续进行漏洞（后门）检测，还部署应用安全网关实施多维度攻击特征识别与精准拦截，形成覆盖模型生命周期的立体化安全防护体系。

除了技术层面的保障，企业还需要关注AI技术的法律与合规性。不同国家和地区的隐私保护法、数据安全法等法规，对于AI模型的安全性提出了具体要求。因此，AI安全治理还需要与政策和法规紧密结合，确保合规性与安全性同步提升。

5.3 数据与隐私问题

随着AI技术的广泛应用，AI模型的驱动力——数据，显得愈加重要。然而，企业在实践中往往面临数据孤岛问题，即不同部门、业务单元甚至合作伙伴之间的数据不能互通或共享。这不仅限制了数据的价值，也影响了AI系统的整体性能。在许多行业中，数据孤岛问题依然较为突出，缺乏高效的数据共享机制和统一的数据平台。

解决这一问题的一个关键方法是通过建设数据共享平台和数据标准化机制，实现跨部门、跨企业的数据共享。通过构建统一的数据架构，企业能够打破数据孤岛，整合分散在不同系统和数据库中的信息，形成完整的数据链条。当前的产品级解决方案包括华为云数据湖探索DLI等，可以提供一站式的数据汇聚、流处理、批处理、交互式分析等能力。这为AI-Native技术提供了更多、更高质量的数据来源，从而提升AI模型的训练效果和决策精准度。

此外，数据安全与隐私保护也是AI-Native技术应用中的难题。AI系统在处理大量用户数据、企业数据时，往往涉及个人隐私、商业机密等敏感信息。如果数据在传输或存储过程中遭遇泄露，将可能导致严重的安全事故，甚至引发法律纠纷。因此，如何保障数据的安全性、隐私性，成为AI技术部署的重要考量。涉及的数据安全技术挑战包括但不限于：数据内容合规性审查、敏感信息与隐私防护、知识库安全保护、数据中毒检测等。对此，华为云DataArts Studio通过对数据进行分类分级、加密存储、权限控制实现多层保障，为客户数据提供全链路全生命周期的安全保护。

在全球范围内，数据隐私法律和政策不断加强，如欧洲的GDPR（通用数据保护条例）和中国的《个人信息保护法》等，都对企业在数据处理中的行为提出了严格要求。为了合规运营，企业需要确保其AI系统在数据采集、处理、存储及使用过程中，能够满足相关法规的要求，同时还需采取数据加密、匿名化处理、访问控制等技术手段，加强数据安全保障。

5.4 异构、多代际硬件的高效协同使用问题

在AI-Native系统的实际部署中，异构、多代际硬件的高效协同已成为关键挑战之一。当前，大模型训练与推理依赖的算力硬件呈现多元化趋势，如英伟达（NVIDIA）的H100/H200 GPU、华为昇腾（Ascend）NPU加速卡等，各家厂商的架构设计、计算指令集、显存配置及通信协议存在显著差异。与此同时，硬件迭代速度极快，企业往往在数年内积累多代设备（如V100/A100/H100），导致集群中存在算力不均衡、显存容量不一、互联带宽不同等问题。如何在不牺牲效率的前提下，实现跨厂商、跨代际硬件的统一调度与协同计算，成为AI基础设施层的核心难题。

在算力层面，不同硬件对FP16/FP8等精度格式的支持度不同，例如NPU通常针对特定计算类型（如矩阵乘加）优化，而GPU的通用性更强，直接混合使用可能导致计算效率下降或精度损失，需通过编译层抽象或运行时动态调度实现算力标准化。在通信层面，跨厂商设备的互联依赖PCIe/RDMA/InfiniBand等通用协议，但高性能计算场景仍需专用链路（如NVSwitch或昇腾HCCS），若集群异构程度高，需引入跨架构通信库或代理层来优化数据交换效率，避免成为分布式训练的瓶颈。

解决这一问题需要软件栈的深度协同优化。一方面，需构建硬件抽象层，使算法工程师无需感知底层硬件差异；另一方面，需设计动态资源分配策略，例如将计算密集型任务（如矩阵运算）调度至NPU，而将逻辑控制或高精度计算交由GPU处理。此外，可结合模型并行技术，按硬件能力切分计算图——如将大模型的注意力层部署在H100上，而FFN层运行在A100上，通过梯度同步与流水线并行掩盖异构带来的延迟。长远来看，行业需推动开放标准（如OpenXLA、ONNX Runtime）的落地，减少生态碎片化，但在此之前，企业仍需通过定制化调度器与中间件，最大化异构集群的利用率。

5.5 模型能力评价体系构建问题

AI-Native系统构建过程中同样面临AI模型的选择难题，主要挑战包括：

- 多样化应用场景的影响：AI-Native技术的应用非常广泛，包括自然语言处理、计算机视觉、推荐系统等。不同应用领域对技术的要求各异，例如在医疗、金融、法律等行业，AI技术的评估标准和关键指标差别较大。因此，统一的评估体系无法涵盖所有场景的特殊需求，评估标准需要针对不同应用场景进行定制。
- 行业特定需求差异化大：每个行业对AI技术的期望和需求不同。在医疗领域，可能更注重技术的准确性和合规性；在金融领域，则强调安全性和透明性；而在创意行业，创造力和创新性可能是更重要的评估因素。这种行业特性意味着，AI-Native技术的评估体系必须根据实际应用背景进行调整，而不能依赖一套固定的标准。
- 主观与客观评估指标的平衡：与传统技术评估不同，AI-Native技术的评估往往涉及更多的主观因素，如创造力、情境适应性、用户体验等，这些是难以量化的。例如，在艺术创作或内容生成中，AI的“创新性”难以用传统的客观指标来衡量。如何平衡主观性和客观性的评估标准，成为AI-Native技术评估中的一大挑战。

5.6 大模型幻觉问题的治理与突破

生成式AI的核心挑战之一在于大模型幻觉（Hallucination），即模型生成的信息与事实、逻辑或用户输入的上下文不一致，表现为事实性幻觉和忠实性幻觉。这种幻觉现象对AI的实际应用带来了诸多挑战和潜在风险。在医疗领域，模型可能提供错误的诊断和治疗建议，危及患者生命；在法律咨询中，它可能引用虚构的法律条文和案例，导致法律风险；在新闻领域，大语言模型可能生成虚假新闻事件，扰乱信息传播秩序，误导公众认知。此外，幻觉问题还会降低用户对模型生成内容的信任，阻碍AI技术在关键领域的广泛应用。

为解决大模型幻觉问题，研究人员已提出多种方法。提升训练数据质量是基础，需确保数据的准确性和时效性，避免数据偏见和错误信息。改进训练过程也很关键，如结合强化学习与人类反馈的混合训练方式，可有效限制模型生成的随机性，提高响应的准确性。此外，优化模型架构，采用双向建模等方式，有助于更好地理解上下文信息。检索增强生成（RAG）技术通过让模型在回复前参考给定的可信文本，确保回复内容的真实性，也被证明是有效的方法之一。

未来的研究方向可能包括：进一步探索更高效的模型训练策略，以更好地平衡模型的创造力和准确性；开发更先进的上下文理解机制，使模型能够更精准地把握用户意图和语境信息；以及构建更强大的事实核查系统，通过交叉验证的方式实时检测和纠正模型生成的错误信息，从而提高模型的可靠性和可信度。

5.7 多Agent协同与自治挑战

随着AI-Native系统从单一模型调用向多智能体（Multi-Agent）协同演进，如何设计、管理与协调具备自主性的Agent集群成为新的核心技术挑战。Agent通过感知、规划、执行和工具使用来完成复杂任务，但其动态性和交互性也引入了在单体模型中不存在的复杂性问题。

- **动态环境下的规划与决策可靠性问题：**Agent的核心能力在于其根据目标进行规划与决策。然而，在开放、动态的真实世界环境中，Agent的规划链条可能因信息不完整、环境突变或自身推理错误而失效。长链条的复杂任务规划中，微小的错误或偏差可能会被逐步放大，导致最终结果严重偏离预期。例如，一个负责市场分析的Agent可能在数据解读阶段发生细微误解，进而导致后续的预测和策略建议Agent做出完全错误的决策。确保Agent在长周期、多步骤任务中规划的一致性和鲁棒性，是其在关键业务场景中落地的前提。
- **多智能体协同的通信与竞争难题：**在Multi-Agent系统中，多个智能体需要通过通信和协作来完成共同目标。这带来了诸多挑战：首先，高效的通信协议亟待建立，Agent之间需要以机器可解析、语义无歧义的方式进行信息交换，避免“沟通误会”。其次，如何设计有效的协同机制，既避免Agent陷入无休止的辩论或循环，又能激发群体智能，实现“1+1>2”的效果。此外，当多个Agent目标不一致或资源有限时，可能产生竞争关系，系统需要具备冲突检测与消解能力，防止因内耗导致系统整体效能下降。
- **工具使用的精确性与安全性风险：**Agent通过调用外部工具（如API、数据库、执行器）来扩展能力边界。然而，工具的使用引入了新的风险层面。在精确性上，Agent需要准确理解工具的功能、输入输出格式，任何参数错误或调用时序问题都可能导致任务失败。在安全性上，赋予Agent工具使用权限等同于授予其操作现实世界的能力。一个未经充分验证的Agent行为，可能错误调用删除数据的API，或向外部系统发送恶意指令，造成难以挽回的损失。因此，必须建立严格的工具调用授权、审计和“急停”机制，确保Agent的行为在安全边界内。
- **身份、记忆与长期一致性的维持：**一个成熟的Agent应具备持续学习和与用户交互的能力，这要求其拥有持久的记忆和稳定的身份。技术挑战在于如何高效、结构化的存储和检索海量的交互历史，并使Agent能够基于过往经验进行学习与自我演进。同时，如何保证Agent在长期运行过程中，其行为模式、价值观与设定初衷保持一致性，避免发生“性格漂移”或被恶意引导，是关乎系统可信度的深层次问题。

为解决上述挑战，业界正在探索一系列技术路径，例如采用分层控制的架构（管理者Agent协调工作者Agent）、为Agent设定明确的行为宪法（Constitutional AI）、在沙箱环境中进行高风险操作模拟、以及建立贯穿始终的可观测性框架，对Agent的决策树、工具调用链和内部状态进行实时监控与记录。这些努力旨在使Agent系统在保持高度自治的同时，兼具可靠性、安全性与协同效率。

06

AI-Native 技术的未来展望

AI-Native技术已进入快速发展期，但仍面临算法效率、环境交互、算力瓶颈、人机协作和伦理治理等多重挑战。这些挑战既是技术演进的关键障碍，也是推动下一代AI系统突破的重要契机。随着新型算法架构、计算范式和人机交互模式的不断涌现，AI-Native技术正朝着更高效、更智能、更可信的方向发展。以下从五个关键维度探讨未来技术演进路径：

1) 算法架构的范式突破：从Transformer到更高效的智能架构

当前以Transformer架构为主体的大模型虽然在自然语言处理等领域取得突破，但其计算资源消耗大、难以高效处理时序依赖与动态环境等固有缺陷逐渐显现。例如，Transformer的自注意力机制在长序列处理中面临计算复杂度与内存占用的指数级增长，导致模型训练成本居高不下。此外，其对静态数据的依赖与对物理世界动态交互的弱感知能力，限制了AI系统在复杂现实场景中的适应性。

未来，新型高效算法通过架构创新有望突破这一局限：Mamba架构通过状态空间模型（SSM）实现线性复杂度长序列建模，显著降低计算资源消耗；SNN（脉冲神经网络）模拟生物神经元动态脉冲特性，在边缘计算场景展现超低功耗优势；类脑计算借鉴大脑的稀疏编码与事件驱动机制，可构建更适应开放环境的通用智能体；具身智能（Embodied Intelligence）通过多模态感知-动作闭环，赋予AI物理世界因果推理能力。这些技术将共同推动AI从“数据驱动”迈向“认知驱动”，而图灵奖得主Yann LeCun提出的世界模型（World Model）通过自监督学习构建环境动态表征，或将成为突破LLM物理规律建模瓶颈的关键。类似地，这些前沿技术的持续探索共同勾勒出AI从感知向认知跃迁的技术图谱，其交叉演进或将成为实现AGI的核心推力。

2) 环境交互的范式跃迁：走向物理世界是AGI形成的必由之路

当前AI系统虽在数字空间中表现出强大的模式识别与内容生成能力，但其对物理规律的认知缺失与环境交互的抽象化处理，仍是实现通用人工智能的核心瓶颈。例如，大语言模型虽能精准描述重力现象，却无法让机器人绕过现实障碍；生成式AI可创作逼真图像，却难以预测推倒积木塔的精确物理轨迹。这种“数字智能”与“物理智能”的割裂，使得AI在开放环境中的适应性、因果推理与持续学习能力受限。

未来，AI必须跨越数字符号与物理实体的鸿沟，通过具身感知与交互构建对现实世界的深层理解——具身智能体将通过多模态传感器实时捕捉环境状态，在物理约束下学习行动策略；世界模型则通过模拟物理动态，为智能体提供预测与规划能力。斯坦福李飞飞团队提出的“空间智能”正推动AI从二维感知向三维交互演进，使智能体不仅能识别物体，更能理解其物理属性与互动逻辑。与此同时，神经符号计算将符号系统的推理能力与神经网络的感知优势结合，使AI在操作物体时既能遵循物理定律，又能进行任务分解与逻辑判断。这一演进并非简单增强现有模型，而是构建能感知、推理并干预物理世界的智能系统，使AI从封闭数据空间走向开放环境，最终在机器人、自动驾驶、工业自动化等领域形成能力闭环。只有完成从“数据智能”到“物理智能”的跃迁，AI才能真正理解并改变现实世界，而这正是实现AGI不可绕过的终极路径。

3) 算力架构的重构：从通用芯片到任务定制化与量子协同

当前GPU和NPU等通用算力架构虽在AI训练与推理中占据主导地位，但其“一刀切”的设计模式难以适配AI应用的多样化需求。例如，GPU在通用并行计算中表现优异，但面对生成式AI的长序列推理任务时，其内存带宽与延迟瓶颈导

致能效比下降；NPU虽针对特定算子优化，却在多模态任务中面临架构灵活性不足的挑战。此外，硬件迭代速度与算法演进的脱节（如Transformer架构的演进速度远超芯片架构更新周期）进一步加剧了算力资源的碎片化。

未来，任务定制化芯片将通过架构创新打破这些限制——Groq的LPU以流水线化计算单元和大规格内存带宽设计，专为生成式AI推理任务优化，使长序列输出效率提升10倍以上；Sohu芯片则通过将Transformer架构“蚀刻”至硬件层，实现算子级能效突破。更深远的演进在于量子计算与AI的深度融合，量子比特的叠加与纠缠特性将为分子模拟、组合优化等复杂问题提供指数级算力跃迁。

4) 人机协同的进化：从工具辅助到共生共创

当前AI系统在重复性任务中展现出强大能力，但在创造性、伦理判断与复杂决策领域仍显不足。例如，AI在艺术创作中难以突破风格模板化，在战略规划中缺乏对人类价值观的动态适配。同时，多Agent协作系统的标准化缺失也制约了跨领域任务的高效协同——不同Agent的接口协议、知识表示方式差异导致协作成本居高不下。

未来，人机协同将迈向“共生共创”新范式——通过模型上下文协议（MCP）与Agent互联协议（A2A）的标准化，实现跨模态知识的无缝融合与多Agent任务的动态编排。例如，医疗领域中，AI Agent可实时分析影像数据并生成诊断建议，人类医生则基于临床经验进行伦理判断与决策优化。在科研领域，AI将作为“超级助手”辅助人类完成复杂建模与实验设计，例如通过自监督学习构建材料科学的因果模型，加速新材料的发现。这种“人机共智”模式将重新定义生产力边界，使AI从“替代工具”进化为“能力增强器”。人机协同将成为未来工作和生活的新常态，通过合理分配任务和发挥各自优势，共同推动社会的进步和发展，创造更加美好的未来。

5) 伦理治理的范式升级：从被动合规到主动价值对齐

当前AI系统的伦理风险已从技术层面扩展至社会层面，例如算法偏见导致的决策歧视、数据隐私泄露引发的信任危机，以及自动驾驶等场景中的道德困境。这些问题的根源在于AI系统缺乏对人类价值观的动态适配能力，其决策逻辑仍基于静态数据与算法规则，难以应对复杂的社会伦理情境。

未来，AI伦理治理将从被动合规升级为主动价值对齐——通过因果推理增强的世界模型，AI系统将具备对物理规律与社会规范的深层理解，例如在医疗诊断中结合医学伦理准则进行决策优化。可解释性AI（XAI）技术的突破将进一步提升算法透明度，使AI的决策过程可追溯、可验证。此外，去中心化的伦理治理框架将通过区块链与联邦学习技术，实现数据隐私保护与多方协作的平衡。例如，医疗AI可在联邦学习框架下训练模型，同时通过零知识证明技术确保患者隐私不被泄露。这些创新将推动AI从“技术工具”向“社会基础设施”转型，使其真正成为人类价值观的数字延伸。

通过上述多维度的演进，AI-Native技术将在算法、环境交互、算力、人机协同与伦理治理层面实现系统性突破，最终构建出高效、智能、可信且可持续的下一代AI生态。这一进程不仅将重塑产业边界，更将重新定义人类与技术的关系，为智能时代的文明进步奠定基础。

后记

本白皮书的发布离不开创原会的鼎力支持，创原会以“聚云上领航者，论AI原生之道”为宗旨，汇聚了一群耕耘在技术一线，立志以技术改变世界的技术管理者、技术专家、学者等，通过创原会的平台洞悉技术前沿发展、交流行业落地实践，共创技术与业务融合的无限可能。

创原会是于2020年由华为云、中国信通院、CNCF三方共同发起成立，致力于成为连接技术、产业与实践的桥梁，在数字经济时代助力企业实现数智跃迁。回顾过去几年创原会发布的技术共识，持续走在产业前沿，引领产业发展：

- 2021年提出“云原生2.0”：带来“资源高效、应用敏捷、万物互联、极致体验、数据融合、业务智能、安全可信”七大新价值；
- 2022年提出“以云原生思维践行云原生”：总结云原生2.0十大新范式“泛在新范式、计算新范式、网络新范式、调度新范式、应用新范式、数据新范式、智能新范式、安全新范式、万物互联新范式、行业使能新范式”；
- 2023年提出“深度云化，成就云原生企业”：通过“分布式云基础设施”实现“上好云”，通过“应用现代化、数智融合”实现“用好云”，通过“精益IT治理、确定性运维、企业级安全、FinOps”实现“管好云”；
- 2024年提出“云原生 x AI，开启数智化跃迁”：实现“架构跃迁、算力跃迁、存储跃迁、数智跃迁、应用开发跃迁、媒体技术跃迁、安全体系跃迁、行业跃迁”八大跃迁；
- 2025年提出“智能进化，全面拥抱AI-Native”：包含“AI-Native基础设施、AI-Native云服务、行业AI大模型、行业AI应用”。

技术的发展日新月异，有创原会这样的一个平台，通过开放交流，推动技术真正解决企业遇到的问题挑战，带来新的技术红利。未来，在新的AI-Native技术方向，相信创原会将继续发挥产业引领的作用。

感谢创原会及所有会员对本白皮书的支持和帮助！

华为技术有限公司
深圳龙岗区坂田华为基地
电话：+86 755 28780808
邮编：518129
www.huawei.com

商标声明
HUAWEI，HUAWEI，是华为技术有限公司商标或者注册商标，在本手册中以及本手册描述的产品中，出现的其它商标，产品名称，服务名称以及公司名称，由其各自的所有人拥有。

免责声明
本文档可能含有预测信息，包括但不限于有关未来的财务、运营、产品系列、新技术等信息。由于实践中存在很多不确定因素，可能导致实际结果与预测信息有很大的差别。因此，本文档信息仅供参考，不构成任何要约或承诺，华为不对您在本文档基础上做出的任何行为承担责任。华为可能不经通知修改上述信息，恕不另行通知。

版权所有 © 华为技术有限公司 2025。保留一切权利。
非经华为技术有限公司书面同意，任何单位和个人不得擅自摘抄、复制本手册内容的部分或全部，并不得以任何形式传播。