未来网络发展大会
FUTURE NETWORK
DEVELOPMENT CONFERENCE
NANJING · CHINA / 中国 · 南京

未来网络技术发展系列白皮书（2025）

DeepSeek行业大模型算力网加速应用生态白皮书

第九届未来网络发展大会组委会

2025年8月

版权声明

编写说明

主要编写单位：

紫金山实验室、江苏省未来网络集团

主要编写人员：

侯聪、张晨、张玉军、高新平、徐政、孙婵娟、芮美芳、杨荣、陈晓波、陈刚、赵芷晴

特别鸣谢（无先后顺序）：

苏交科集团、笑领科技、贵州师范大学、江宁数据局

新华三、浪潮信息、中兴通讯

南京智能计算中心、火山引擎、算力互联、安徽提尔液冷科技

天数智芯、沐曦集成电路、燧原科技、昆仑芯科技、海光信息、寒武纪、壁仞科技

前　言

自 OpenAI 问世后，各路大模型如雨后春笋般涌现，它们基于互联网上爬到的数据进行训练，要花费成千上万张 GPU 资源才能训练出来，这些大模型可以陪人闲聊、回答问题甚至求解方程，但是它无法知道的企业流水线的工艺制造方式、学校对学生的个性培养计划、医院为老人的病症诊疗方案。这些大模型我们称其为“通用大模型”，它知道的很多很杂、但不深不准。如果要让 AI 真正服务于千行百业，需要的是把“通用大模型”与行业数据充分结合，再通过算力加工成“行业大模型”。“行业大模型”的发展，需要迈过三座大山，一是模型部署太贵，企业要部署一套聪明的大模型动辄要大几百万上千万，二是数据流通不畅，企业的生产数据通过互联网传输既不安全也不高效，三是算力使用不便，国家建设了大量的公共算力却未充分被企业所知所用。

DeepSeek 为“行业大模型”的征程开了个好头，它既足够聪明又开源免费，将大模型成本直降到 0。模型的问题解决了，数据流通和算力使用的问题又该如何解决？今年春节前后，DS 云部署和一体机一时间蔚然成风：云部署基于互联网提供了轻量化的 DS 服务，但面对 B 端市场存在着“数据传不出，网络运不动，算力信不过”的众多约束；一体机基于局域网部署本地化的 DS 设备，可有效解决云部署的上述挑战，但在 B 端落地应用时则面对着“建设成本高、服务性能僵、模型更新慢”等新的挑战。

为实现云部署与一体机两者优势的“兼而得之”，未来网络联合多方合作伙伴，共同打造了基于算力网加速的 DeepSeek 行业大模型边云一体化解决方案，可实现全域的算力按需供给、数据可信流转与模型实时同步，有效破解鱼掌难题：（1）高效安全的数据传输，基于确定性网络调度技术，可保障超过 1000GB 的自动驾驶数据不到 5min 就传输完毕，而传统网络则需要 10 天左右，让时间“不耗在路上”；（2）全局协同的资源调度，基于计算存储网络协同调度技术，可根据全网计算资源的动态调度结果自动匹配相应的数据存储与网络传输资源，实现“货朝店走，路随货通”；（3）案前手边的使用入口，基于调度系统边缘接入技术，可通过一体机或更轻量的小盒子实现一键加速全网通达，作为企业接入东数西算的算力阀、算力表，把“路铺到家门口”。

众所周知，东数西算与全国一体化算力网的最终目标，就是让千行百业像用水和用电一样用算。如何实现这个目标？（1）对于用算方而言：回想一下，在通水网之前我们需要找一口水井、买一个水缸，这就类似于当前云计算基于门户网站下单购买算力资源，当通水网之后我们只需要在屋子里面装一个水龙头、装一块水表，这就类似于未来算力网基于算力阀、算力表动态使用算力资源，需要就用不需要就停，用多少付多少，真正实现“最优匹配、按需启停”；（2）对于供算方而言：在通电网之前，水电、风电、轮机、集中式光伏、分布式光伏都有不同的发电技术与电流特性，但经过并网后都将转换至标准电压大小与统一电流特性，当前云计算中业务逻辑复杂难归一，基

于门户下单购买算力资源的方式，就好比用户指定用何种发电技术为自己发电，未来算力网中大语言模型业务逻辑可统一抽象为 Token 输入输出，基于算力阀、算力表动态使用算力资源，用户无需指定厂商、架构、型号，进而拉动国产算力的充分消纳，真正实现“精准计量、效用付费”。

《DeepSeek 行业大模型算力网加速—应用生态白皮书》（简称白皮书）的编制，得到了来自应用场景方、一体机设备厂商、算力服务商、国产芯片厂商等众多合作伙伴的大力支持，白皮书的发布期间正值全国一体化算力网并网、计量等国家标准技术文件的出台制定，希望能够为国家东数西算与全国一体化算力网提供未来网络实践经验，未来能够赋能每个企业都能够拥有自己专属的企业大模型、让各个行业都能发展出领域的行业大模型，走出一条我国特色的 AI 发展与应用路线。

目　　录

一、现状与挑战

1.1 DeepSeek

自 2024 年 3 月到 2025 年 3 月，我国大模型在一年之内先后完成了从技术（2024.3，Kimi 长文本重大升级）—— 产品（2024.5，豆包上线头条/抖音）—— 市场（2025.2，DeepSeek 现象级爆火）的华丽转身，DeepSeek 的“深度慢思考”获得了“全民加速度”。

DeepSeek 深度思考的产品能力与国民出圈的市场热度，引发了 C 端使用 DeepSeek 的风潮。在 DeepSeek 之前，业界并非没有开源大模型，但当时它们要么血统不够纯正（如 Llama 部分开源）、要么智商不够聪明（如 Qwen 开源小参数）、要么情商不够细腻（如 GLM 主要 2B）。DeepSeek 集血统（充分开源）、智商（深度思考）、情商（人文关怀）于一身，一经发布就获得了万千宠爱，加之“东升西落”话题的论战式传播，一时全球震惊。

除了 C 端出圈以外，DeepSeek 开源更为深远的意义在于 B 端。在 DeepSeek 之前，私有化部署一套足够聪明的大模型动辄要大几百万上千万，令企业望而却步。而 DeepSeek 的开源将模型成本直降为 0，一时间全行业竞相争用。DeepSeek 自身作为通用大模型，它知道的虽多、却杂，但当企业落地应用 DeepSeek 并将其与自身管理生产经营数据充分结合，就能够让 DeepSeek 掌握的更深、更准。未来的不久，DeepSeek 将不再仅仅是陪人闲聊、回答问题甚至求解方程，

而且它能够知道企业流水线的工艺制造方式、了解学校对学生的个性培养计划、知悉医院为老人的病症诊疗方案，让大模型不仅能够飞入千家万户，更能走进千行百业。

目前，DeepSeek 已经在众多行业得到了应用，在落地过程中企业也已经逐步从对 671B 的盲目追风恢复到 32B/70B 的理性落地，各类智能体也渐渐走进了企业的办公与管理流程。虽然如此，但客观上而言 DeepSeek 目前仍存在很多先天缺陷，其中最大问题之一就是 DeepSeek-V3/R1 目前仍然不具备多模态能力，“只会听不会看”使其在很多业务生产场景有心而无力。未来，随着 DeepSeek-R2 的发布，这块短板一定会得以弥补，并在实际落地中扮演更多关键角色。

1.2 行业大模型

从“通用大模型”到“行业大模型”，并非一步之遥，更非一日之功。从演进路线来看，“通用大模型”首先需要结合各个企业自身的数据变为“企业大模型”，然后汇集多个企业的数据/智能方能变为“行业大模型”。

“企业大模型”目前已经在众多企业起步，其使用细致来看又可分为推理和微调两种方式。推理的架构是“大模型+知识库”，其本质可理解为“查字典”，虽然稍显机械但已经能够解决企业日常办公、管理中的很多问题。微调的架构是“大模型*数据集”，其本质可理解为“背字典”，其好处是在背字典的过程中可能会组合并涌现出新的知识以便“举一反三”，但其处理过程会消耗大量资源增加使用成

本。对于企业而言，推理和微调并不矛盾，通常是先基于开源大模型旁挂知识库进行推理，以供日常使用，当有效数据积累到一定程度时就可以进行一次微调，微调生成新的大模型后，可以对推理所用模型进行更新，以此往复加以时日，企业就能够真正拥有自己的“专属大模型”。

“行业大模型”目前更多地还处于研究阶段，其潜在方式可分为以下几种：（1）将一个行业内部多家典型企业的数据统一汇集，然后通过算力直接训练得到行业大模型，这种方式虽然直截了当，但跨企业汇集数据的难度较大，在现实层面可行性受限；（2）把一个行业里面的多个企业大模型汇聚在一起，通过一个行业大模型的入口来进行问题的分流和答案的整合，这种方式架构上与 MoE 有所类似但存在分权分域问题，目前技术路线仍在摸索之中；（3）把一个行业里面的多个企业大模型汇聚在一起，通过群体协作方式相互启发增智从而得到行业大模型，这种方式较为新颖但同时具有无限潜力，是未来应重点探索的技术路线。

纵观行业大模型的演进路线，我们处在企业大模型阶段，正从“推理”走向“微调”。不过，无论是推理还是微调，当前在企业落地中都仍面临着不小的挑战与问题。虽然模型成本的大山已经被 DeepSeek 移除，但数据流通不畅、算力使用不便仍影响着 DeepSeek 在企业中的规模应用。

1.3 算力网加速

企业部署 DeepSeek 的方式，主要可分为云部署和一体机两种，春节期间各大公有云争先恐后适配 DeepSeek，力求通过更好的资源弹性与更强的安全防护，让 DeepSeek 问答不再“服务超时”，不过节后落地时却发现企业更青睐于本地部署，于是各路 IT 厂商纷纷发布 DeepSeek 一体机，主打一站式交付和傻瓜式使用。

两者相比而言，云部署基于互联网提供了轻量化的 DeepSeek 服务，但面对 B 端市场存在着“数据传不出，网络运不动，算力信不过”的众多约束，即：企业关键的业务数据不敢随意地传到互联网上，即使敢传出去在互联网上传的也比较慢，而且传到云上还担心公有云窥视窃取自身的业务数据。一体机基于局域网部署本地化的 DeepSeek 设备，可有效解决云部署的上述挑战，但在 B 端落地应用时则面对着“建设成本高、服务性能僵、模型更新慢”等新的挑战，即：企业不仅需要为一体机的采购投入不菲的项目成本而且后续还面对着日常的用电与运维开销，采购回来的机器在使用规模与性能上具有明确的限制而无法灵活调节，机器上虽然预装了当前版本的模型文件但一旦模型升级就束手无策。

实际上，云部署和一体机是一个硬币的两面，云部署的优势就是一体机的缺点，而云部署的劣势恰恰是一体机的优点。当下，云部署和一体机可以说仍然处在对立面上，未能融会贯通。但对于企业而言，事情本不应是非此即彼的选择。

如何破解这种僵局？算力网即是理想的解决方案。2025 年 4 月，

《全国一体化算力网监测调度平台建设指南》公开征求意见，对算力网的内涵进有如下阐述：“算力网可通过专用网络实现入企、入园、入校、入户，为用户提供最优匹配、按需启停、精准计量、效用付费的算力资源供给能力，使用户获得即开即用的用算体验”。

如果将一体机看作算力网入企、入园、入校、入户的载体，那么我们就可以在一体机上加装一种叫做“算力阀”和“算力表”的能力：当一体机本地资源充足时优先使用本地资源，当访问突发而本地资源不足时即发生“需求溢出”，再通过算力网动态调配云端算力加以灵活补充，当访问下降时云端算力自动释放直至为0。“算力阀”和“算力表”就正如水网中的“水龙头”、“水表”，用户只需要拧开阀门就能连接到全国范围算力网上的算力资源，也不必关注这些算力资源的归属、架构、位置，从而真正实现“用水用电一样用算”。

二、算力网加速解决方案

2.1 方案定位

基于上述现状与挑战，未来网络团队设计并研发了面向DeepSeek 行业大模型的算力网加速解决方案。基于国家东数西算安全新总线的广域确定性网络传输能力，连接用户本地与包括国家东数西算八大枢纽节点在内的全网算力资源，突破网络传输安全性、可靠性和速率瓶颈。当用户本地算力不足以支撑大模型业务时，可利用算力网调度平台，将本地任务请求动态溢出至性价比最高的云端算力。

未来网络致力于构建便捷、高效、可信的全网算力分销渠道，集成异属、异构、异地算力资源并感知其状态，支持通过软件服务、型号硬件、一体机集成等多种方式实现用户接入，并为用户提供 PaaS 层数据、算力、模型的一体化任务式调度和按需服务能力，为用户呈现**极简（任务一键加速）**、**柔性（访问无级变速）**、**安全（数据可信流转）**的使用体验，以及任务运行状态可观、访问效果评级可感、账单透明真实可信的服务闭环。

方案通过构建“前端轻量化交互+后端分布式计算”的新型算力网入口形态，使得用户无需了解技术参数细节，只需聚焦大模型应用的业务目标，即可一键获取最优性价比云端资源，进而突破单机性能瓶颈，拓展大模型训推业务范围、显著缩短高并服务延迟与模型更新时间，有效解决由本地资源不足和互联网性能限制引起的服务受限和体验降级问题。产品采用 Serverless 容器按需启停、精准计量和效用

付费特色服务体系，相比纯本地算力方式，可节省用户推理成本达50%。同时，方案可有效盘活云端国产算力资源，赋能算力供应方拓展分销渠道、扩大用户规模，进而实现算力中心的资源消纳。

2.2 总体架构

2.2.1 功能架构

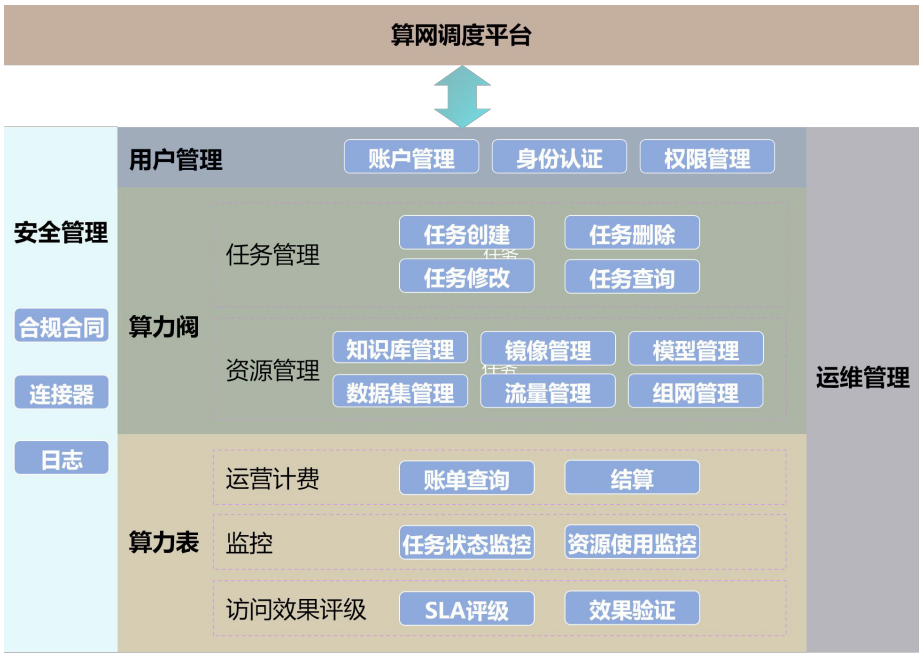

图 2-1 功能架构

方案总体功能架构如图 2-1 所示。其中，算力网调度平台支持任务式的算网协同调度能力，可实现数据、模型跨广域的自由高效流通。基于算力网调度平台，本白皮书将重点阐述用户接入调度平台进行大模型业务加速所需的功能架构，主要包括以下五个模块：

（1）用户管理。负责用户的账户管理、身份认证与权限管理，具体包括创建、信息维护、状态管理、密码重置、注销等账户全生命周期管理，用户名密码、多因素认证（MFA）、单点登录等身份认证

功能，以及定义 RBAC 等细粒度访问控制策略的权限管理。

（2）**算力阀**。负责增、删、改、查等云端任务的操作管理，以及任务相关资源管理，包括对接知识库生成提示词、推调业务镜像的纳管与算力资源适配、模型全生命周期管理、微调数据集管理、业务流量管理及网络组网管理。

（3）**算力表**。负责账单查询、在线结算等运营计费功能，任务状态监控、资源状态上报信息接收与展示等监控功能，以及访问效果评级功能，主要对微调效果及推理首字时延等服务质量进行评级。

（4）**运维管理**。负责产品统一运维管控与效能优化，提升业务效率及连续性，包含基础设施层的采控与对接，平台层的版本配置管理、流程管理、监控告警、智能运维，应用层的全栈自动化远程运维作业、可视化综合大屏及分类分级视图等模块。

（5）**安全管理**。负责用户私有知识库、数据集、模型与镜像的安全出域，在数据分级分类与用户合规基础上，构建可信数据空间（连接器），基于分布式架构与区块链智能合约互操作框架，运用数据主权保障与隐私增强技术，实现各要素在授权范围内的可信流通。

2.2.2 组网架构

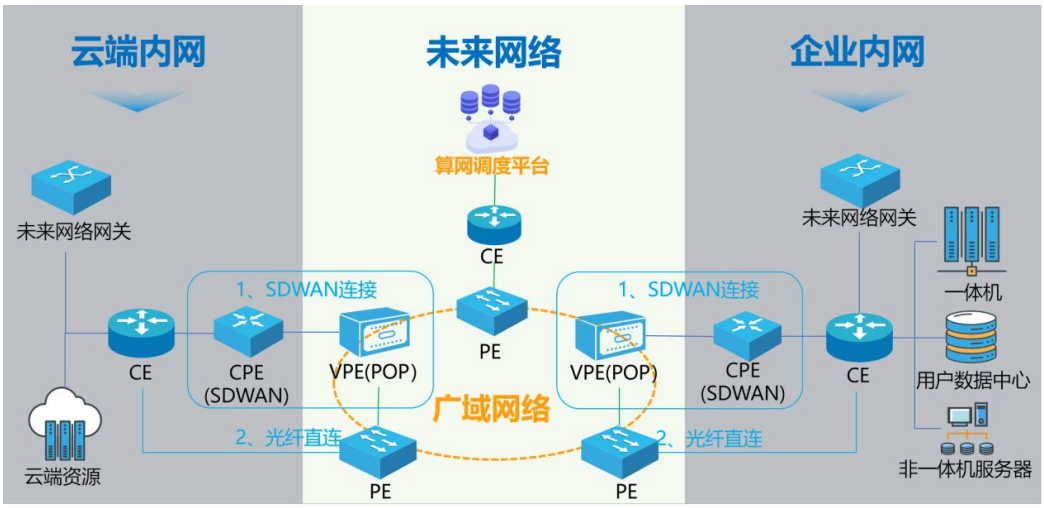

图 2-2 组网架构

用户企业内网以及云端算力内网均可以通过隧道与光纤/专线等方式接入广域网络，进而与算力网调度平台互通，实现边云一体的算力网加速调度，两种连接方式的组网架构分别如图 2-2 所示。

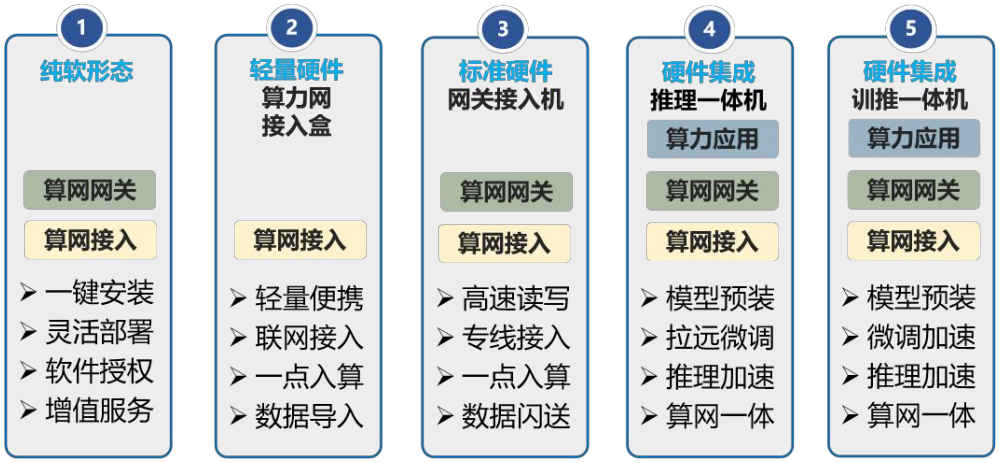

图 2-3 部署形态

业务的部署形态如图 2-3 所示，支持用户以纯软件、轻量硬件、标准硬件、硬件集成等多种形态接入加速平台：

- 纯软件。支持一键安装，能够灵活部署在容器、服务器等载体上，用户通过购买软件授权获取算力网加速增值服务；

- **轻量硬件。**算力网接入盒具有轻量便携、联网接入、一点入算和数据导入等特点，支持以 SDWAN 通过互联网接入平台，并实现本地业务与云端业务的负载均衡；
- **标准硬件。**网关接入机覆盖处理任务、数据、流量等网关接入功能，实现高速读写、光纤专线接入、数据快递，以及基于接入服务的全域负载均衡等能力；
- **硬件集成（推理一体机）。**具有预装模型，能够拉远微调拓展本地不支持的微调业务，并对本地推理业务进行加速；
- **硬件集成（训推一体机）。**为用户提供模型预装、微调加速、推理加速、算网一体的大模型云边协同推调业务。

2.2.3 部署方案

推理应用层典型部署方案分为以下四种：

（1）如图 2-4 所示，推理部署应用业务平台（包含本地部署和云端增值部署）、推理会话平台部署在 CPU 服务器上；大模型训推一体机作为本地算力节点，包含推理模型和推理引擎镜像，并部署推理实例；知识库、状态库和模型管理系统独立部署。

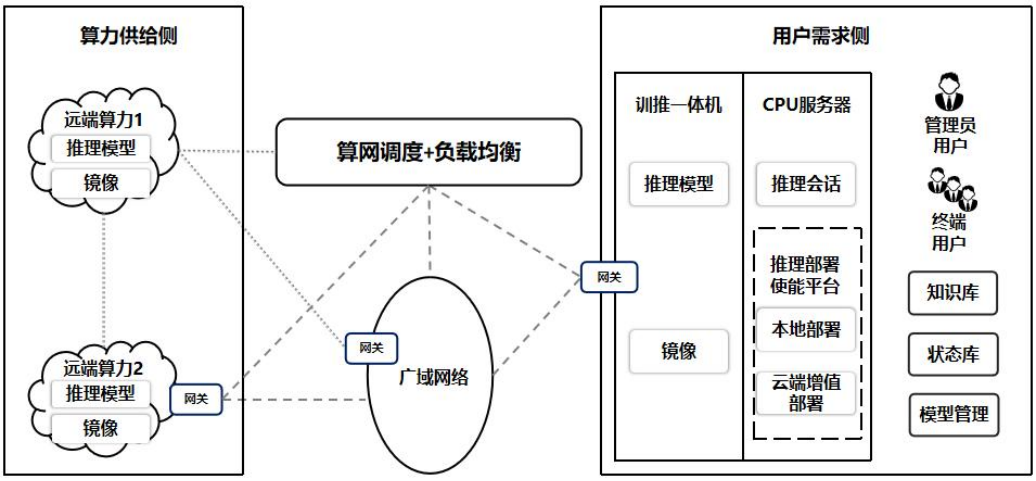

图 2-4 推理应用层典型部署方案 1

（2）如图 2-5 所示，推理部署应用业务平台（云端增值部署）、推理会话平台部署在 CPU 服务器上；用户侧无本地算力，仅能选择远端算力进行模型部署和推理服务；知识库、状态库和模型管理系统独立部署。

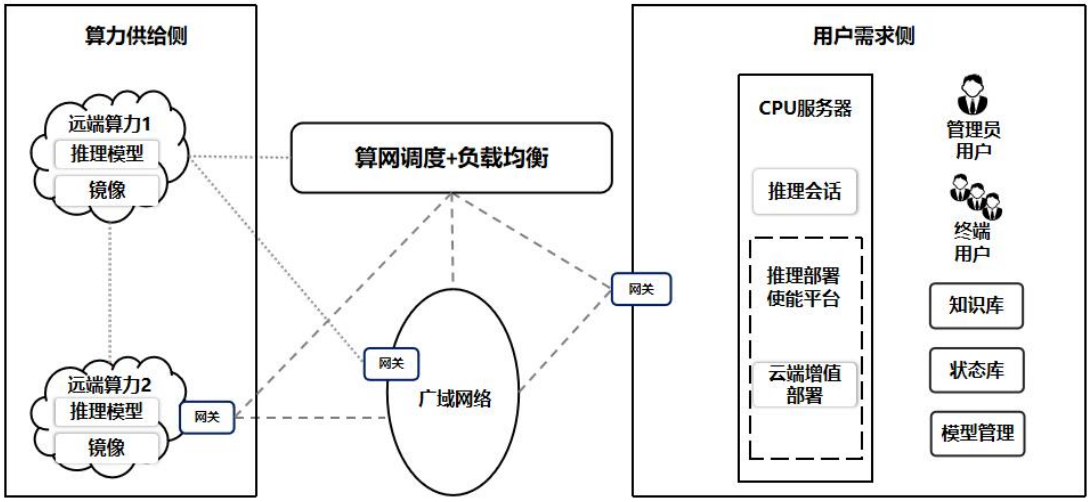

图 2-5 推理应用层典型部署方案 2

（3）如图 2-6 所示，推理部署应用业务平台（包含本地部署和云端增值部署）、推理会话平台、推理模型和推理引擎镜像均部署在大模型训推一体机上；知识库、状态库和模型管理系统独立部署。

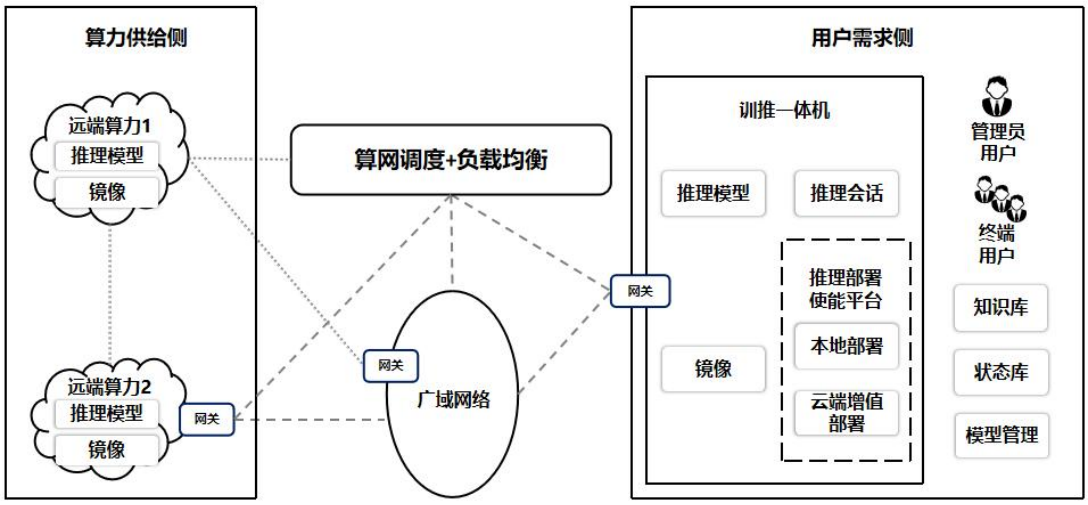

图 2-6 推理应用层典型部署方案 3

（4）如图 2-7 所示，推理部署应用业务平台（本地部署）、推理会话平台、推理模型和推理引擎镜像均部署在大模型训推一体机上；推理部署应用业务平台（云端增值部署）部署在独立的 CPU 服务器上，管理员用户需要进行远端推理部署时，使用独立的云端增值部署界面下发任务；知识库、状态库和模型管理系统独立部署。

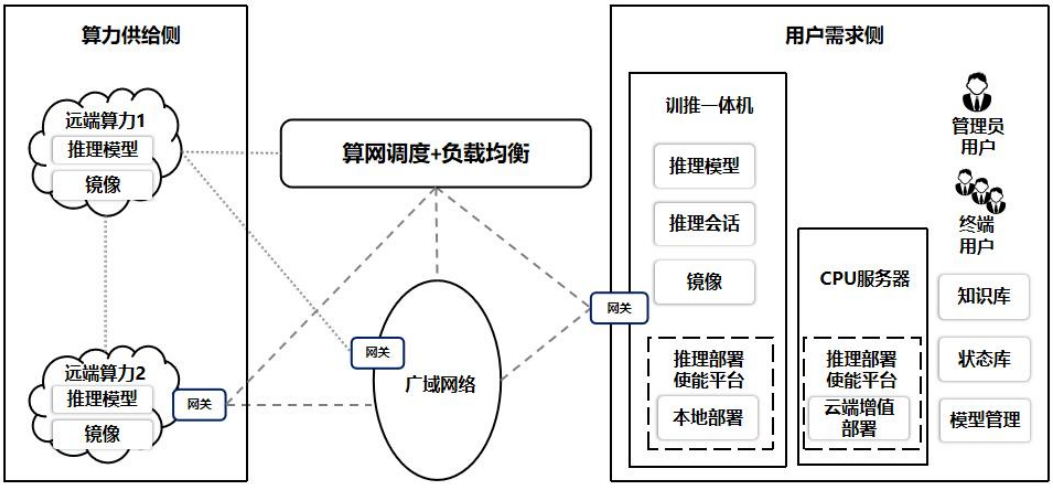

图 2-7 推理应用层典型部署方案 4

微调应用层典型部署方案也可分为以下四种：

（1）如图 2-8 所示，微调部署应用业务平台（包含本地部署和云端增值部署）部署在 CPU 服务器上；大模型训推一体机作为本地

算力节点，包含微调前源模型文件和微调镜像文件；用户数据中心和模型管理系统独立部署。

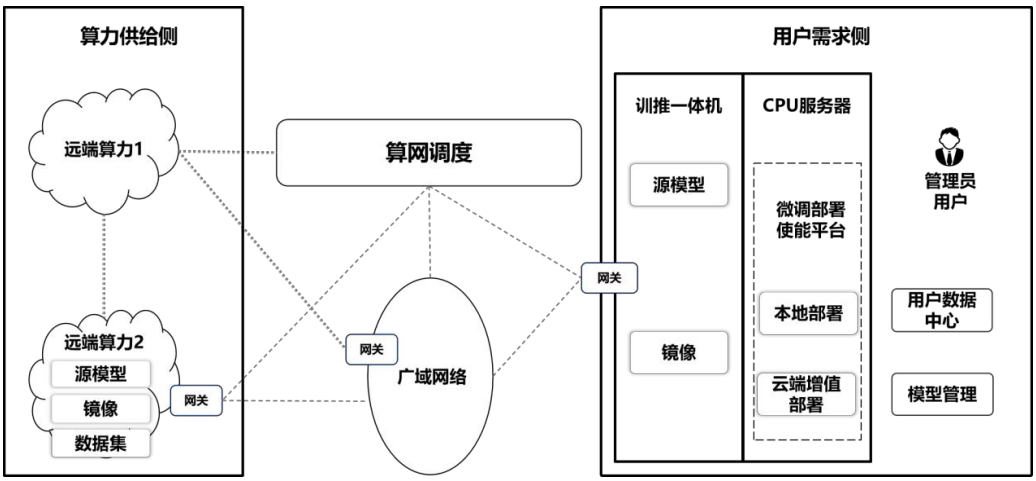

图 2-8 微调应用层典型部署方案 1

（2）如图 2-9 所示，微调部署应用业务平台（云端增值部署）部署在 CPU 服务器上；用户侧无本地算力，仅能选择远端算力进行微调；用户数据中心和模型管理系统独立部署。

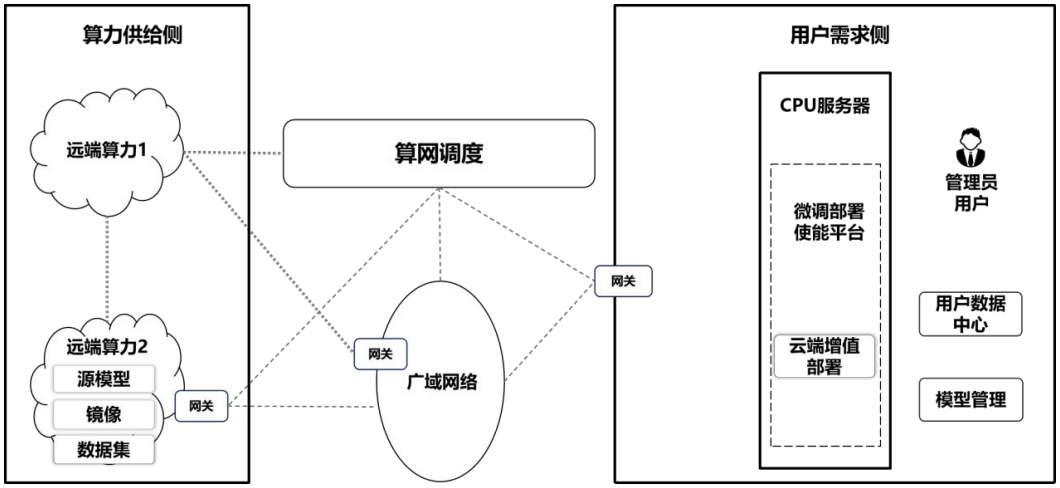

图 2-9 微调应用层典型部署方案 2

（3）如图 2-10 所示，微调部署应用业务平台（包含本地部署和云端增值部署）、微调前源模型文件和微调镜像文件均部署在大模型训推一体机上；用户数据中心和模型管理系统独立部署。

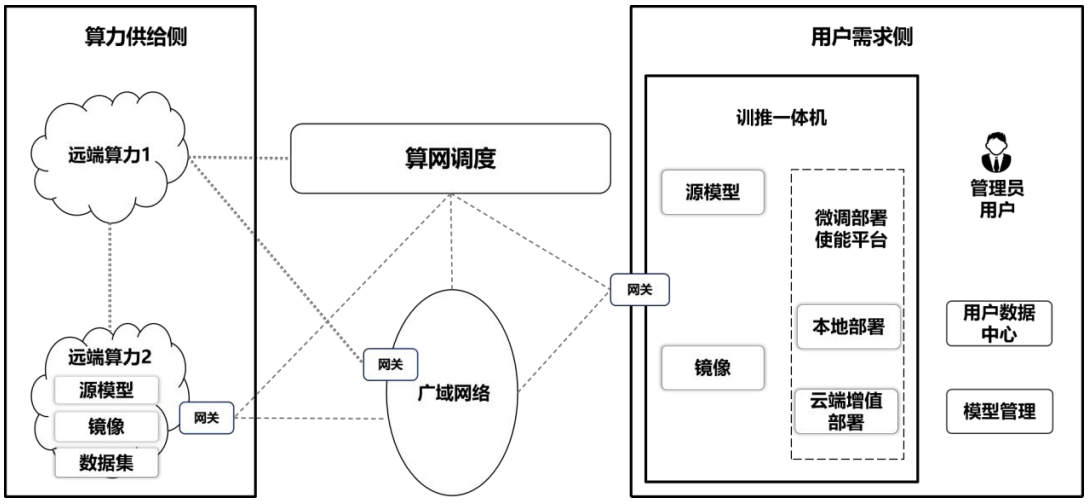

图 2-10 微调应用层典型部署方案 3

（4）如图 2-11 所示，微调部署应用业务平台（本地部署）、微调前源模型文件和微调镜像文件均部署在大模型训推一体机上；微调部署应用业务平台（云端增值部署）部署在独立的 CPU 服务器上，管理员用户需要进行远端微调部署时，使用独立的云端增值部署界面下发任务；用户数据中心和模型管理系统独立部署。

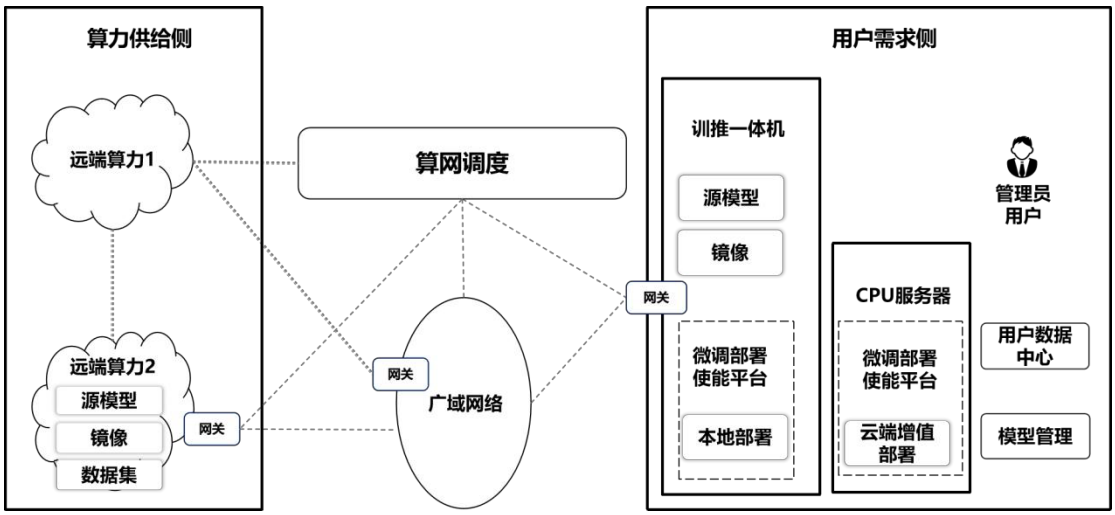

图 2-11 微调应用层典型部署方案 4

2.3 业务流程

2.3.1 推理加速

算力网调度与加速的推理业务包括两个阶段：推理任务协同调度与负载均衡推理服务。在调度阶段，管理员用户通过主动查询监控、终端用户反馈、系统自动提示等渠道获悉加速需求，发起资源调度请求，随后调度平台选定靠近用户的边缘算力或远端高性价比算力，以无服务器（Serverless）占位的方式将资源绑定，并同步信息至负载均衡模块。在推理服务阶段，由终端用户发起推理会话，负载均衡模块根据最小路径等均衡策略，将业务流量分流至本地或远端算力。

在调度阶段，若存在多个可用资源，将进行无服务器占位，直至推理服务阶段，负载均衡模块将终端业务流量分发至占位节点后，调度平台再继续完成推理镜像在该节点的实际部署，即“算随网动”模式。

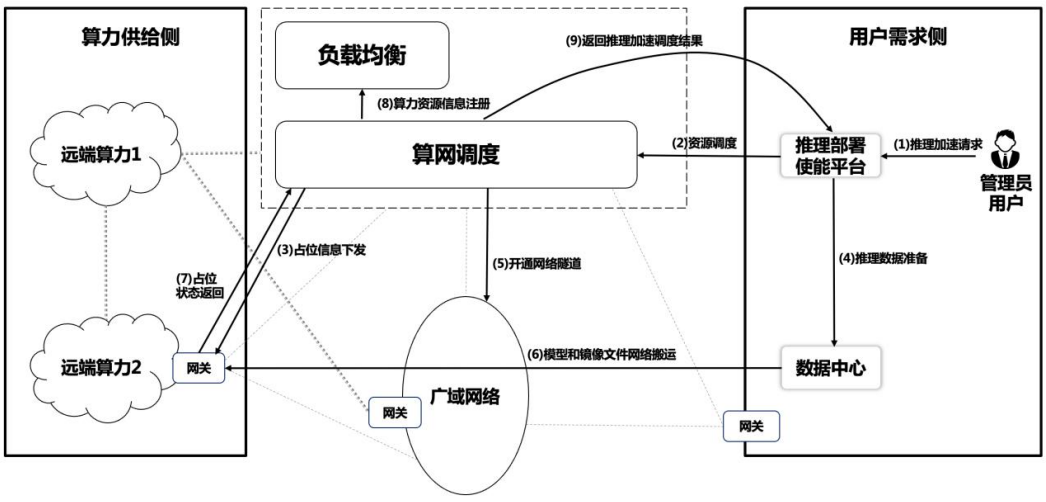

图 2-12 推理调度阶段业务流程图

推理调度阶段的业务流程如图 2-12 所示，具体如下：

（1）管理员用户在推理应用业务平台发起云边协同的推理调度

请求，其中携带当前任务相关的推理模型、费用、时延、任务模式等要求信息。

（2）推理应用业务平台同步调度请求至算力网调度平台。调度平台根据大模型算网需求图谱、用户画像以及本次请求中携带的用户输入要求信息，自动补齐算力厂家、型号、位置、卡数、网络时延等算网规格要求，完成算力、网络与存储资源的云边协同调度。

（3）调度平台以 Serverless 方式绑定调度方案中的资源信息。

（4）推理应用业务平台发起推理数据准备工作。

（5）调度平台开通广域网络传输隧道。

（6）广域网络传输隧道将模型文件和推理引擎镜像文件等同步至已调度的远端算力节点。

（7）远端算力节点返回当前占位状态。

（8）算力网调度平台向负载均衡系统同步推理调度结果，包括算力节点路由信息，以及资源状态、推理部署情况等监控信息。

（9）算力网调度平台向推理应用业务平台返回推理调度结果。

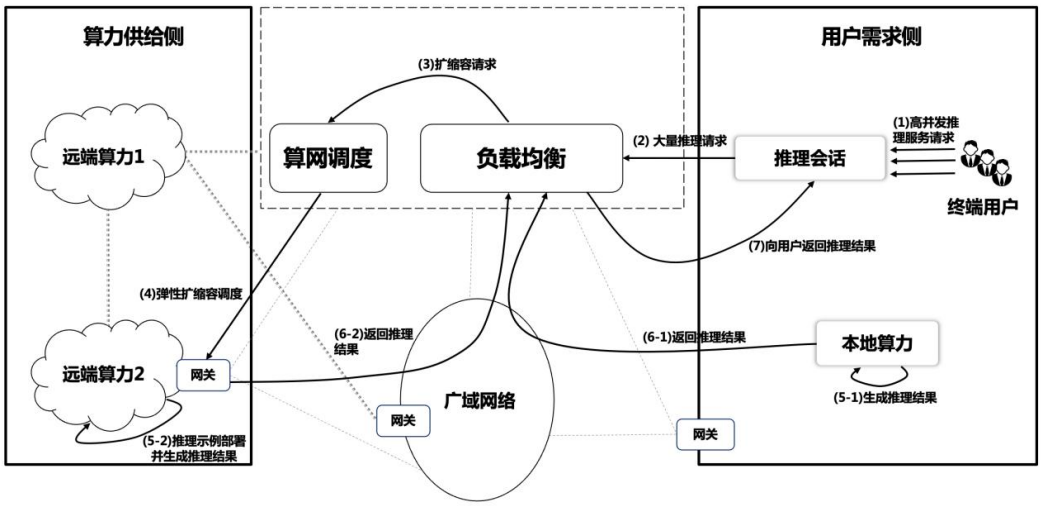

图 2-13　推理服务阶段业务流程图

推理服务阶段的业务流程如图 2-13 所示，具体如下：

（1）终端用户向推理会话平台发起高并发推理服务请求。

（2）负载均衡系统收到大量推理请求，并为业务流量提供包含本地算力节点和远端算力节点的全局路由信息。

（3）负载均衡系统向算力网调度平台发起算力资源弹性扩缩容请求。

（4）算力网调度平台根据远端算力的资源利用率实时下发扩缩容调度指令。

（5）包含以下两个并行步骤：

（5-1）本地算力节点生成推理服务结果，继续步骤（6-1）。

（5-2）等待扩容的推理实例完成部署后，远端算力节点生成推理服务结果，继续步骤（6-2）。

（6）包含以下两个并行步骤：

（6-1）本地算力节点向负载均衡系统返回推理服务结果。

（6-2）远端算力节点向负载均衡系统返回推理服务结果。

（7）负载均衡系统向推理会话平台返回推理服务结果。

2.3.2 微调加速

微调业务仅涉及管理员用户。用户本地算力不足，无法微调或微调排队时间过长时，管理员用户发起微调部署任务，将微调任务部署在远端算力节点，然后在该节点执行微调业务，以缩短任务排队时间、拓展本地业务范围，即“网随算动”模式。

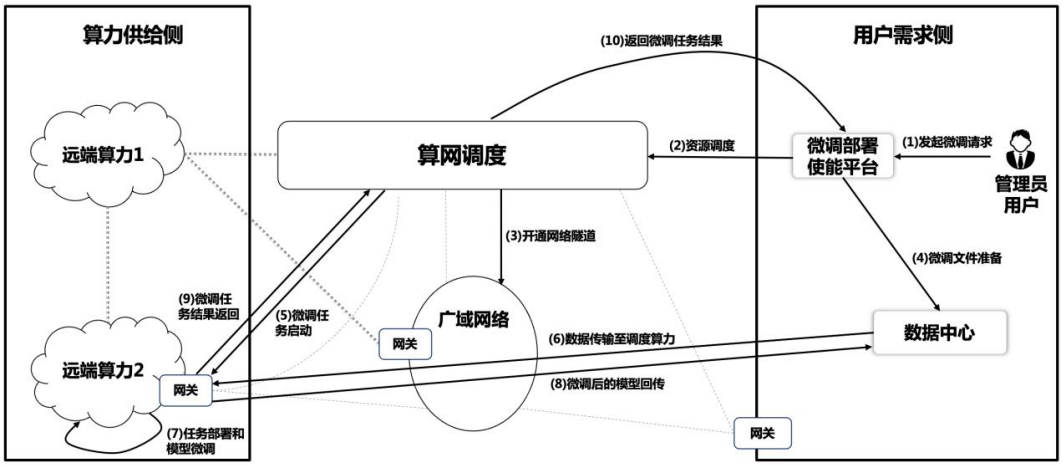

图 2-14 微调业务流程图

微调调度与加速业务流程如图 2-14 所示，具体如下：

（1）管理员用户在微调部署应用业务平台发起微调算网调度和部署请求，其中携带微调模型、任务费用、任务模式、微调完成时间、微调数据集、微调后模型回传路径等信息。

（2）微调部署应用业务平台请求算力网调度平台进行微调远端调度和部署。

（3）算力网调度平台开通广域网络传输隧道。

（4）微调部署应用业务平台发起微调文件准备请求，包括微调源模型文件、微调镜像文件、微调使用的用户数据集等。

（5）算力网调度平台向远端已调度算力下发微调启动指令。

（6）广域网络传输隧道将已准备的微调文件传输至调度的远端算力。

（7）远端算力部署微调任务，进行模型微调。

（8）模型微调完成，回传至数据中心。

（9）远端算力向算力网调度平台返回本次微调任务的结果。

（10）算力网调度平台向微调部署应用业务平台返回本次微调任务的结果，并向用户展示。

2.4 关键能力

2.4.1 极简接入

（1）极简资源配置

由于配置算力网调度与加速任务的用户通常不具备专家知识，难以确定满足 DeepSeek 大模型全尺寸型号推理与微调（推调）业务需求的资源规格。为此，需构建算力网业务需求图谱，旨在免除用户手动配置算力资源信息，实现极简化配置与自动化调度。该图谱通过评测国内外主流算力卡对大模型全链条推调业务的支持能力，明确加速任务所需的算力规格，最终实现异属、异地、异构算力的统一对齐与自动调度，为国产 GPU 分销模式突破及大模型云边协同一键式配置提供坚实的科学依据。

算力网业务需求图谱的构建步骤如下：首先，通过分析大模型推调业务典型场景的计算密集需求（如矩阵运算效率）和通信密集需求（如多卡互联、分布式推调同步延迟），建立典型场景的业务需求模型。其次，在充分调研各厂家支持不同推调业务的显卡型号、显存规格、单卡算力、所需卡数、多卡辅助配置要求、可用卡位置及相应一体机信息的基础上，开展场景化算力度量评测验证。评测核心指标聚焦于使用不同算力卡时的推理首 Token 延迟（TTFT）、每 Token 延迟（TPOT）、每秒查询数（QPS）以及微调完成时间等（详见第 3

章）。最后，构建 DeepSeek 全尺寸模型推调业务的“模型-场景-算力-资源”四维图谱，该图谱将特定场景下的关键参数（如模型参数规模、精度要求、服务 SLA 等）映射为具体的算力需求向量（例如 70B 模型推理需算力≥200 TFLOPS、显存≥80GB），并进一步与实际算力资源信息关联，为大模型云边协同自动化调度提供核心策略输入，显著简化用户操作。

（2）无感交互体验

用户以本地大模型服务为入口，即可无感接入算力网调度与加速平台，无感交互体验由以下三方面能力支撑：

任务自动化创建能力。管理员用户手动创建算力网调度与加速任务时，配置模型、业务类型（如推理、微调）、任务预算（费用）、任务模式（如省心模式、放心模式）、时延要求（SLA）等信息，若用户选择将本次配置保存为任务模板，则在后续任务执行过程中，平台会实时监测算力资源状态，一旦检测到本地算力资源不足或大模型推调任务排队积压，平台将自动触发创建调度与加速任务。此过程对用户透明，实现“无感升级”，有效保障关键业务的连续性与时效性。

资源灵活调度机制。平台默认依据预先构建的 “模型-场景-算力-资源”四维业务需求图谱，调度匹配最优的算力资源，极大简化了用户接入配置和操作流程，显著降低使用门槛。同时，平台提供高级配置入口，供具备专业知识的高级用户根据特定需求，手动指定或精细调整算力资源（如指定算力厂商、GPU 卡型号、集群位置或网络配置等）。

任务可选启动模式。为满足不同管理员的风险偏好和操作习惯，平台支持两种任务启动模式：省心（自动择优）模式下，系统根据四维图谱及当前资源状况，自动选择最优算力资源方案并直接执行，无需用户确认，最大化操作便捷性；放心（用户确认）模式下，系统同样提供推荐的资源方案及其预估费用，但需管理员用户确认方案后任务才会启动，此模式赋予用户最终决策权，提升操作可控性与透明度。

无感接入交互体验设计，使平台既能通过高度自动化服务大多数用户，提升效率与业务连续性，又能为专家用户保留深度控制能力，并通过灵活的交互模式适配不同管理需求，最终实现算力资源调度的高效化、智能化和用户友好化。

2.4.2 柔性访问

（1）负载均衡

通过“标识感知—动态决策—弹性闭环”三位一体的全域智能负载均衡技术，实现业务流量的云边跨域动态分发。构建覆盖资源属性-业务特征-网络状态的多维度算力网标识体系，并基于标识体系利用强化学习与博弈论模型，实现多目标优化的请求分发机制。创新性融合算力网标识策略与动态调度算法，优先保证本地业务运行，在本地资源不足情况下，根据地理位置、QPS 加权等负载均衡策略，辅以多级探针、故障自愈等健康检查增强机制，将大模型推调业务请求分发至云端，实现多业务场景下的精细化流量治理，保障高并发、异构化、跨地域环境下的服务等级协议（SLA）业务闭环。

（2）弹性伸缩

平台具备弹性伸缩能力，根据业务需求、算力状态及用户前期配置，自动发起调度与加速任务。业务并发量过大超过算力水位时，进行弹性扩容；业务并发量过大低于算力水位时，将云端资源释放。通过差异化业务的弹性伸缩策略，打破传统互联网不可被调度现状，实现算力与网络的协同调度及扩缩容，以扩容为例：对于微调业务，调度系统完成云端算力网调度后即启动实例进行扩容，再将业务流量分发至云端；对于推理业务，调度阶段对可用算力资源仅进行无服务器占位操作，并未真正启动实例，在推理服务阶段业务流量被负载均衡分流至云端后，才启动云端实例进行扩容。

（3） 模型适配

基于“模型—算力卡—业务类型—服务协议—镜像文件”的五维映射矩阵，实现镜像文件的多维动态适配与全域统一纳管，从应用层确保云边协同调度与加速的灵活普适性及业务连续性。

推调业务镜像的框架与代码因模型架构、算力硬件等条件差异存在显著分化，例如当用户本地采用 A 卡推理时，其镜像（无论来自 A 卡厂商、算力经销商或自主开发）难以适配云端异属异构算力环境，导致本地业务无法在云端实现无级加速。通过全网镜像统一纳管与多维映射矩阵构建，形成覆盖全量模型、调度异构算力、支持混合任务部署、兼容多种协议的镜像全维度适配能力。该机制可提前完成云端环境镜像适配或生成同等性能替代方案，有效解决跨域协同镜像不兼容问题。

2.4.3 安全流转

云边协同推调业务需要将用户本地的私有知识、数据集、模型、镜像文件上传至云端，为保证这些私有资产能够安全可信地跨域跨空间流通，构建算力网调度与加速**可信数据空间**。采用分布式架构，将数据分散存储在多个节点上，避免集中存储带来的风险和依赖性，通过建立一套信任机制和规范，确保多主体数据的真实性和可信度，使数据提供者及使用者相互信任并顺利开展数据共享与流通。

在数据文件分类分级的基础上，利用数字智能合约技术描述算力提供方、一体机厂商、管理员用户、推理服务用户、调度与加速平台等各个参与方对数据文件、使用方式、使用次数等流通使用行为的预期，并达成共识。通过数字智能合约对数据的共享、流通和使用进行规范约束，确保数据安全合规，建立统一的管理制度、技术标准、业务流程，不与现行法律法规相抵触，同时兼容现有的各类技术标准，允许数据提供者和使用者自由协商并定义使用策略，在满足不同主体需求的同时，提供更加个性化的数据流通解决方案。

通过集成在特定软硬件环境中的数据沙箱、隐私保护计算、多租户隔离等使用控制技术，对使用数据文件的算法、应用进行控制和审计，实现对访问、分析、计算等行为的管控。采用基于联盟链的区块链存证技术，将数据哈希、操作日志上链，确保数据完整性、操作不可篡改，对数据使用阶段的共享数据、存储、使用、销毁等全流程进行日志记录，内置 GDPR、CCPA 等法规合规检查工具，自动生成数据流转审计报告，保证数据流通及使用过程如约执行，结果符合预期。

2.4.4 可观可感

（1）任务运行状态可观

通过动态拓扑仪表盘，实时映射本地设备与云端集群的立体化算力资源及网络状态，绘制推理与微调任务流水线，统计任务并发量、分布及实时状态。以热力图、水位图等形式展示 CPU/GPU 利用率、内存负载与网络流量密度等资源饱和度；以拓扑流量图监控网络节点间吞吐量，以时空矩阵视图呈现 GPU 池分布与弹性伸缩状态。针对推理任务，分析边缘预处理、云端聚合、结果回传时延；针对微调任务，跟踪数据上传、模型回传、资源释放等流程，并监控梯度同步状态、参数更新轨迹及迭代轮次等微调过程关键指标。当 GPU 占用率持续超过预警阈值时，自动触发预测性弹性扩容，显著降低大模型推调业务异常处理耗时。

（2）访问效果评级可感

通过智能可视化看板展示微调与推理业务核心指标，即时、精准、敏捷地评估访问效果，直观把握模型及算力网状态，实现云边协同推调业务的合理评分、精准归因、敏捷优化。针对微调业务，绘制精度-时延曲线、算力能效矩阵，并通过 A/B 测试等方式对比微调前后模型的损失函数收敛轨迹及参数分布变化；针对推理业务，及时评估推理访问的 TTFT、TPOT、QoS 等关键指标，预测性规避 SLA 违约；通过集成 AutoML 技术自动优化引擎，实现微调参数自主寻优与推理 SLA 自保障。

三、算力网资源量化测评

3.1 测评概述

为使用户能够仅关注自身大模型业务需求，而无需关注具体所用的算力资源类型，算力网调度系统需掌握各型号算力资源对DeepSeek 大模型的支持情况，以便优化调度策略提升用户体验。以此为驱动，未来网络团队开展了多维的资源量化测评。传统算力测评的目标，是帮助用户更好地选购算力资源，而本测评旨在为算力网调度提供依据，从而在保证服务质量的前提下，可自动调度在网算力资源为用户提供服务，从而使得用户并不感知所使用的算力归属、架构与位置。

本次测评的算力资源，包括了 H20、L40、V100 等三种英伟达GPU，以及 5 款主流国产智算芯片，在多样化的卡数及并发设定下，对 DeepSeek 大模型的吞吐量、延迟、QPS、Token 生成速率等多维度关键指标展开量化测评，为算力网调度策略的制定与优化提供了丰富详实的性能基线。本节将重点对于其中：输出 Token 吞吐量、每一并发平均吞吐量等吞吐性能结果、首 Token 延迟（TTFT，Time To First Token）、每 Token 输出时间（TPOT，Time Per Output Token）等时延性能结果进行介绍。

3.2 测评环境

本次测评基于 vLLM 框架，针对 DeepSeek 两种模型在 8 款算力

卡上的推理性能进行全面评估。由于各种算力卡对 vLLM 版本及模型精度的支持情况有所不同，本次测评结果不能完全代表算力卡的实际芯片性能，在总体测试条件尽量一致的前提下，记录测评环境和操作的差异性，具体测评环境信息请参考表 3-1。

表 3-1 测评环境说明

参数	**测评配置**	
模型种类	DeepSeek-R1-Distill-Qwen-32B	
	DeepSeek-R1-Distill-Llama-70B	
推理框架	H20	vllm-0.8.x
	L40	
	V100	
	E 卡	
	A 卡	vllm-0.9.x
	D 卡	
	B 卡	vllm-0.6.x
	C 卡	vllm-0.10.x
输入长度	1024	
输出长度	1024	
最大上下文	5000	
精度	H20	bfloat16
	A 卡	
	B 卡	

	C 卡	
	D 卡	
	E 卡	
	V100	float16
	L40	bfloat16、float16

3.3 吞吐测评分析

通过在不同数量的算力卡环境上运行 DeepSeek-R1 的 32B 和 70B 模型，分析输出令牌（Token）吞吐量与每一并发吞吐（输出 Token 吞吐/并发数）的情况，如图 3-1 至图 3-7 所示。

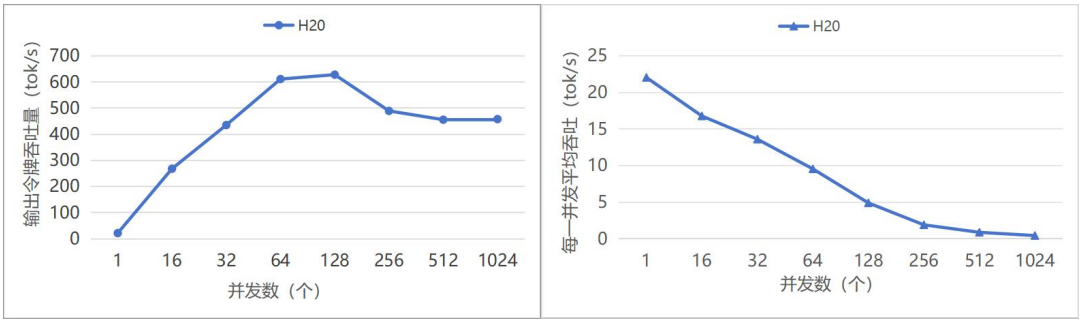

图 3-1　32B 1 卡运行的输出吞吐（左）与每一并发吞吐（右）

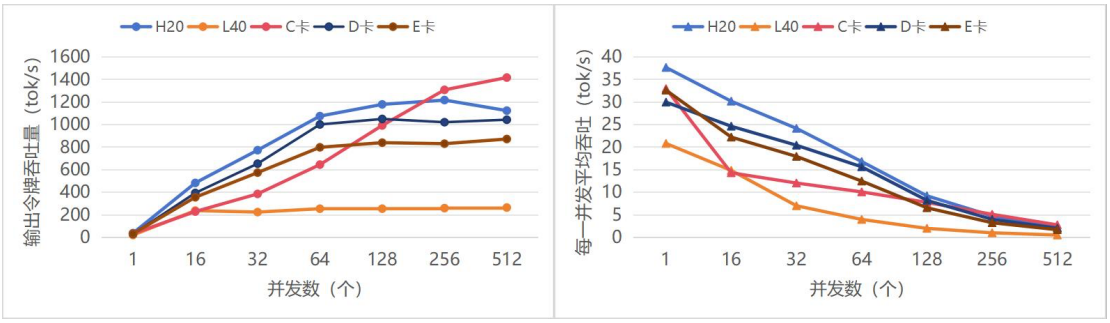

图 3-2　32B 2 卡运行的输出吞吐（左）与每一并发吞吐（右）

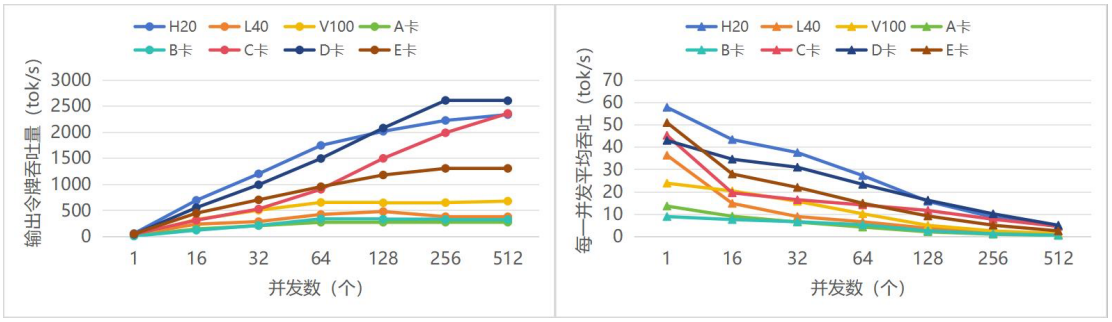

图 3-3　32B 4 卡运行的输出吞吐（左）与每一并发吞吐（右）

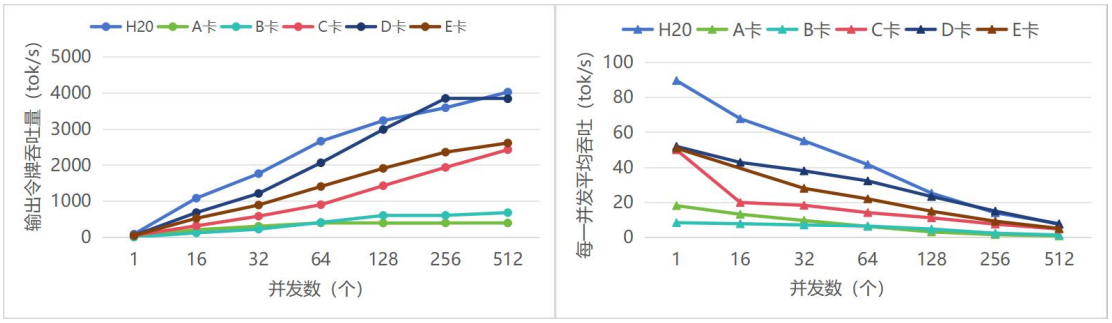

图 3-4　32B 8 卡运行的输出吞吐（左）与每一并发吞吐（右）

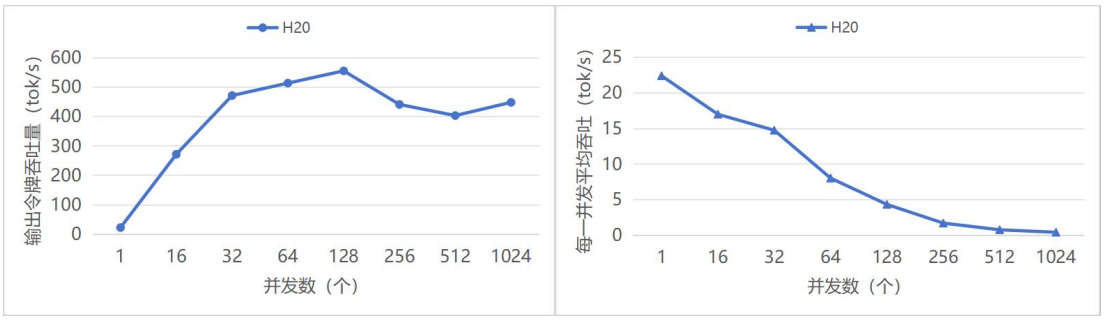

图 3-5　70B 2 卡运行的输出吞吐（左）与每一并发吞吐（右）

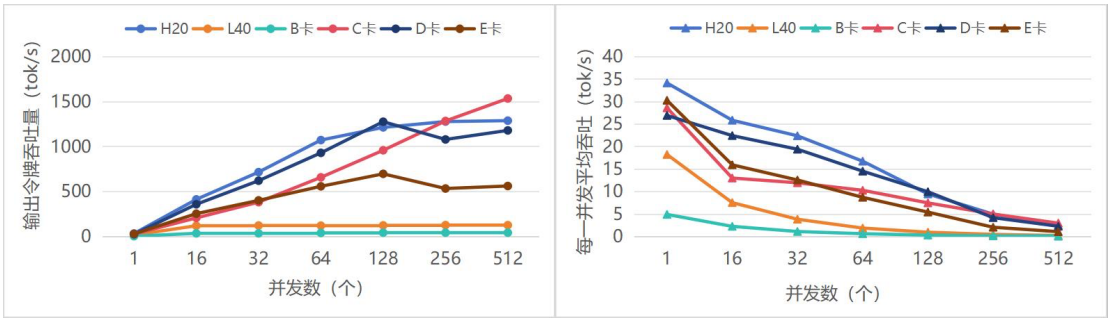

图 3-6　70B 4 卡运行的输出吞吐（左）与每一并发吞吐（右）

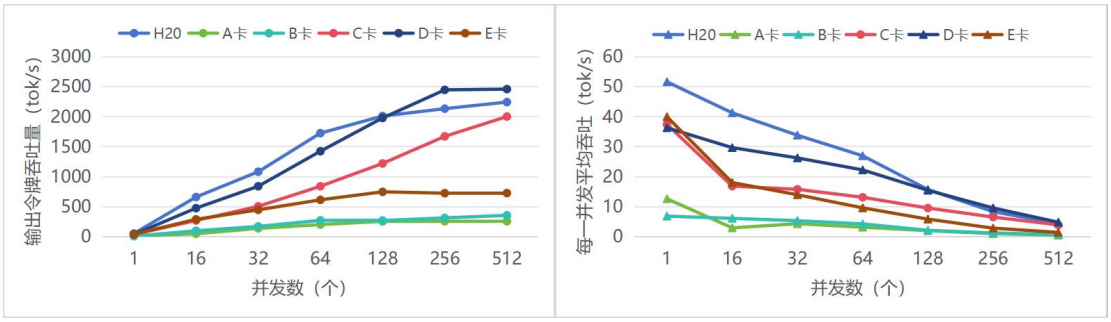

图 3-7　70B 8 卡运行的输出吞吐（左）与每一并发吞吐（右）

从测评结果来看，各种卡型的吞吐表现在 7 种不同测试条件下相对稳定，本次测评总体性能从高到低排序依次为 H20、D 卡、E 卡、C 卡、V100、L40、A 卡与 B 卡，但是因为各种算力卡支持的 vLLM 版本不统一，本次结果仅反应基于当前测评版本的算力卡业务性能。

值得一提的是，本次测评中 L40 性能表现并不理想，运行 32B 模型时性能不如 V100，通过测评 bfloat16、float16（性能影响不明显），以及更改最大显存占用、添加推理加速参数等方式多次核验，L40 的测评结果均不理想。我们认为这种情况一方面是由于 L40 无 NVLink 卡间互联，而 V100 使用了 NVLink 互联；另一种可能的原因是 L40 缺乏专用 Tensor Core 对 float16 及 bfloat16 进行硬件加速，而是通过 FP32 模拟计算，需要额外的 FP32 缓存用于存储中间结果，导致计算效率下降，而 V100 的计算无需此步。

通过分析吞吐数据，我们发现针对在特定卡数下运行特定模型的情况，随着并发数增加，各个卡型的输出 Token 吞吐呈现从快速增长到趋于平稳的走势，每一并发平均吞吐则是从快速下降到趋于平稳。从低并发阶段随并发数敏感变化，到最后趋于平稳，这一过程主要受运行环境的最大吞吐量影响，对应的并发数拐点随卡数的增加而增加，

随模型增大而下降。例如，E 卡在 2 卡运行 32B 模型时，输出 Token 吞吐的并发拐点为 64 并发；在 8 卡运行 32B 模型时，输出 Token 吞吐的并发拐点大于 256 并发；在 8 卡运行 70B 模型时，输出 Token 吞吐的并发拐点为 128 并发。

表 3-2　SLA 限定每一并发吞吐 15 token/s 时的最大并发

算力卡型号	**算力卡数量**	**32B 最大并发**	**70B 最大并发**
H20	1	24	—
	2	80	30
	4	149	80
	8	250	140
L40	2	16	—
	4	15	6
V100	4	33	—
A 卡	4	1 并发已经不满足 SLA	—
	8	9	—
B 卡	4	1 并发已经不满足 SLA	1 并发已经不满足 SLA
	8	1 并发已经不满足 SLA	1 并发已经不满足 SLA
C 卡	2	15	—
	4	42	13

	8	50	40
D 卡	2	66	—
	4	160	60
	8	256	135
E 卡	2	48	—
	4	65	20
	8	127	22

通过分析测评结果，可以推断更多性能指标，为调度策略优化提供坚实的数据基础。例如，在推理 SLA 要求每一并发平均吞吐不低于 15token/s 的情况下，根据每一并发吞吐曲线，估算如表 3-2 所示的各种环境最大支持并发数，进而在并发数超过阈值之前进行提前预判与扩容加速，快速匹配满足并发量与 SLA 需求的合适资源，显著提升用户体验。

3.4 时延测评分析

使用 32B、70B 模型在多样化推理并发量情况下，测评不同算力卡的 TTFT 及 TPOT，如图 3-8 至图 3-21 所示。

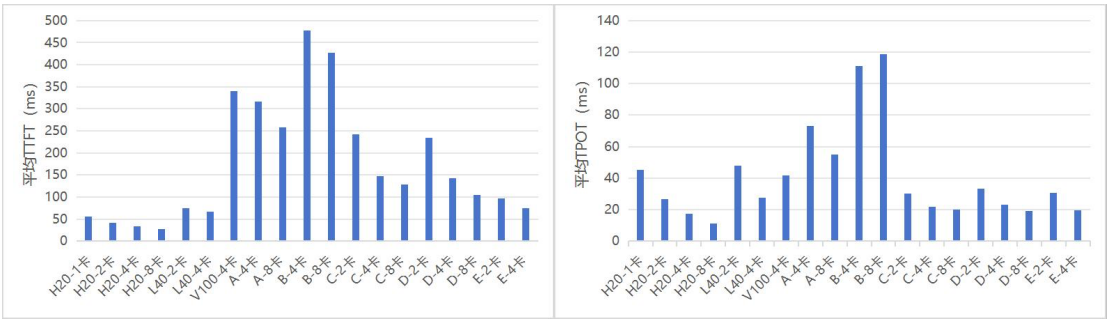

图 3-8　32B 1 并发的平均 TTFT（左）与 TPOT（右）

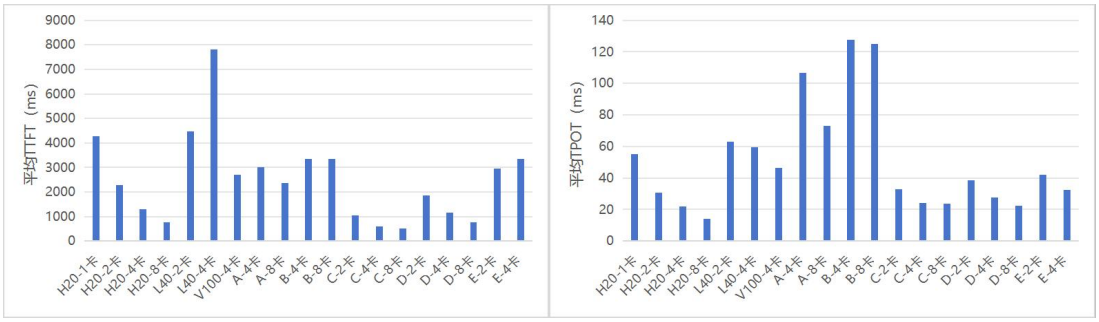

图 3-9　32B 16 并发的平均 TTFT（左）与 TPOT（右）

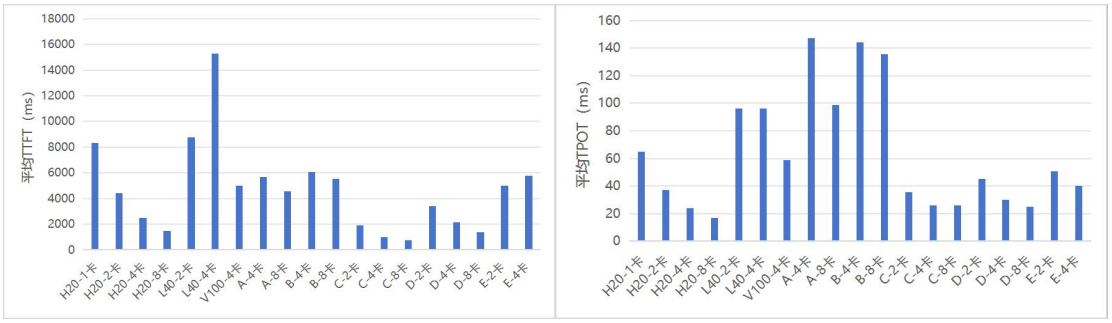

图 3-10　32B 32 并发的平均 TTFT（左）与 TPOT（右）

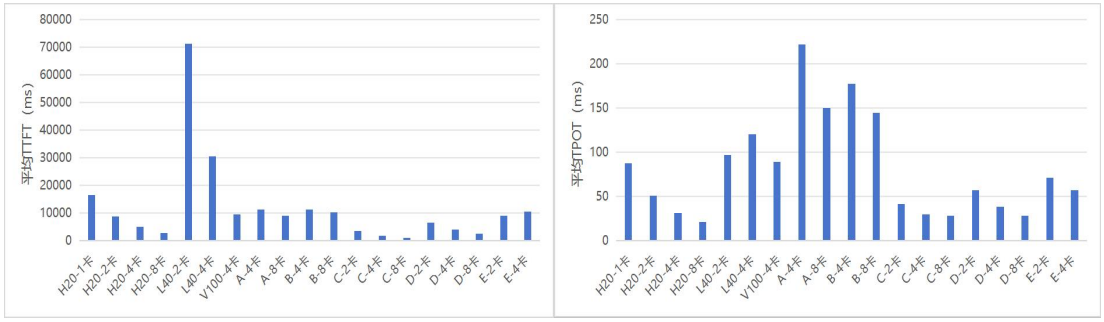

图 3-11　32B 64 并发的平均 TTFT（左）与 TPOT（右）

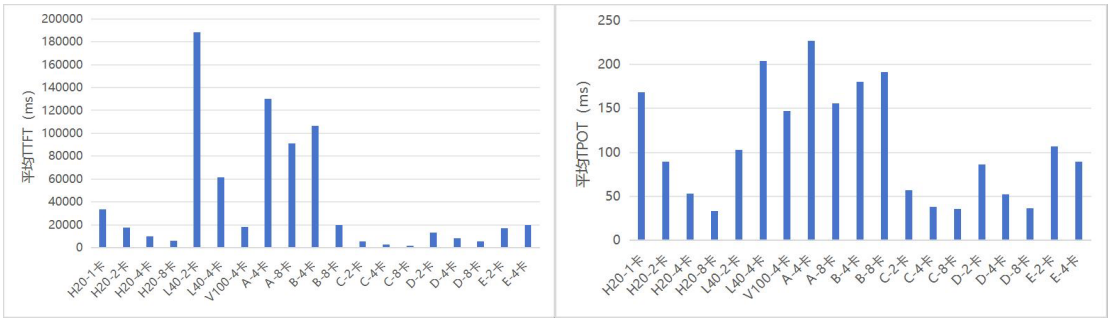

图 3-12　32B 128 并发的平均 TTFT（左）与 TPOT（右）

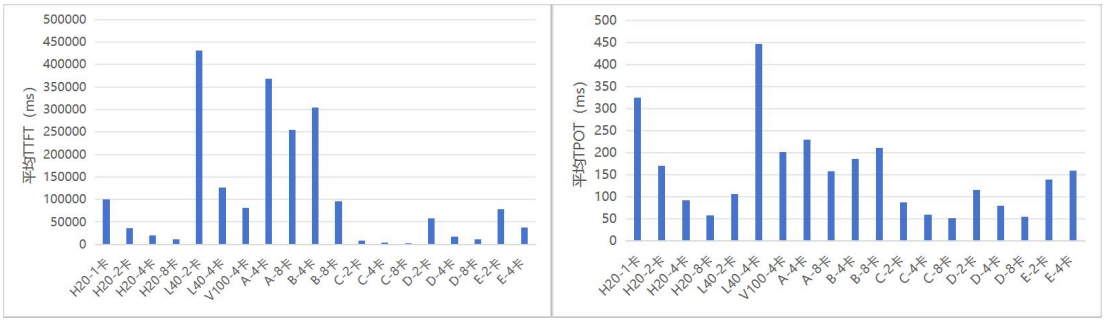

图 3-13　32B 256 并发的平均 TTFT（左）与 TPOT（右）

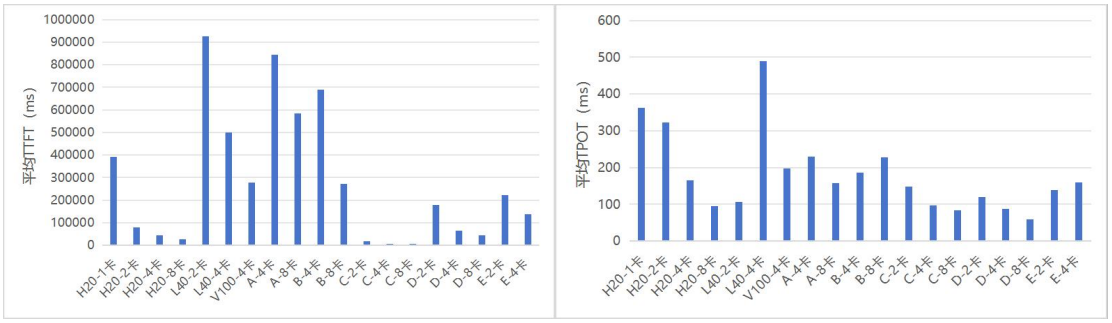

图 3-14　32B 512 并发的平均 TTFT（左）与 TPOT（右）

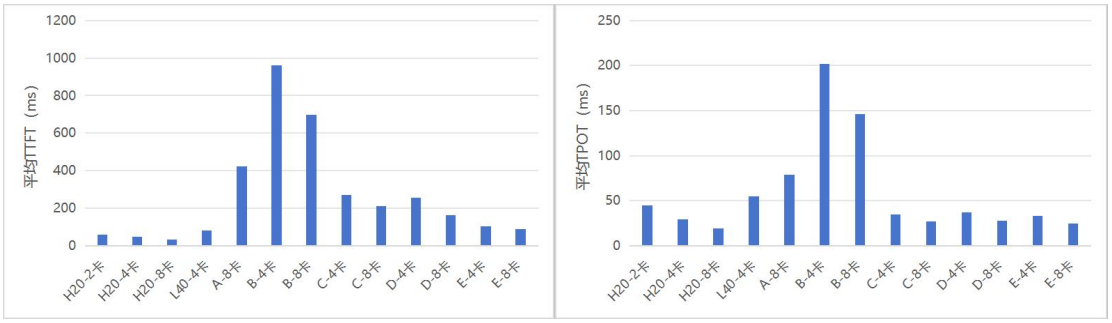

图 3-15　70B 1 并发的平均 TTFT（左）与 TPOT（右）

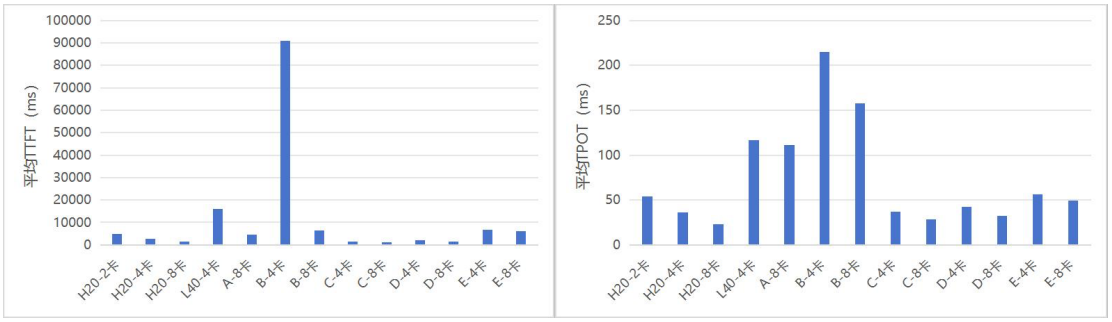

图 3-16　70B 16 并发的平均 TTFT（左）与 TPOT（右）

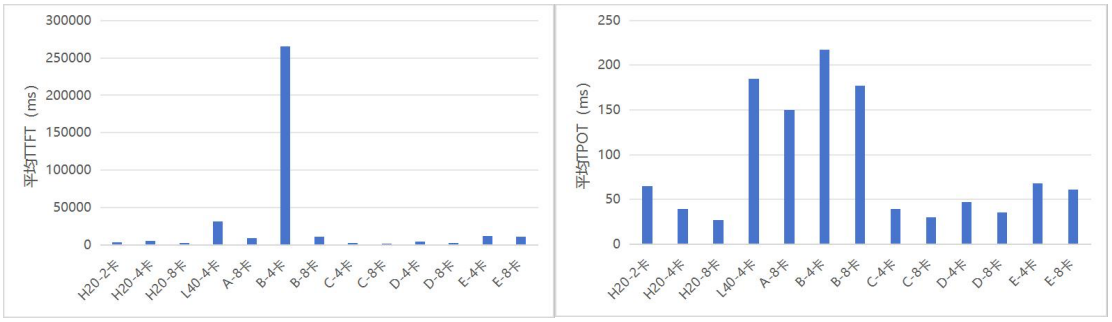

图 3-17　70B 32 并发的平均 TTFT（左）与 TPOT（右）

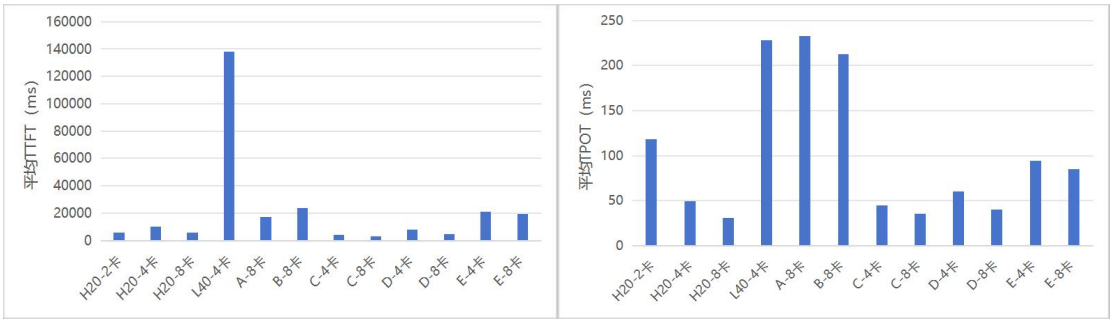

图 3-18　70B 64 并发的平均 TTFT（左）与 TPOT（右）

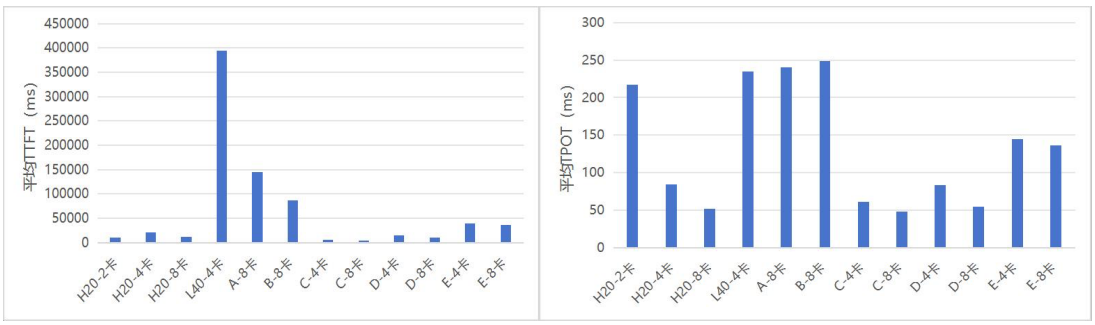

图 3-19　70B 128 并发的平均 TTFT（左）与 TPOT（右）

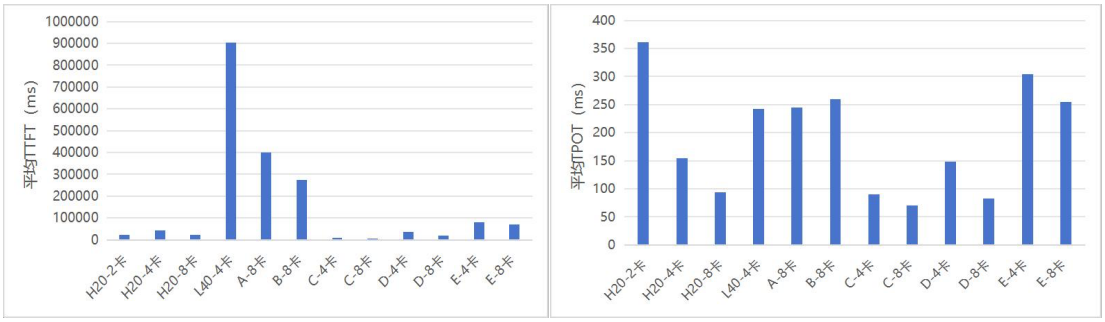

图 3-20　70B 256 并发的平均 TTFT（左）与 TPOT（右）

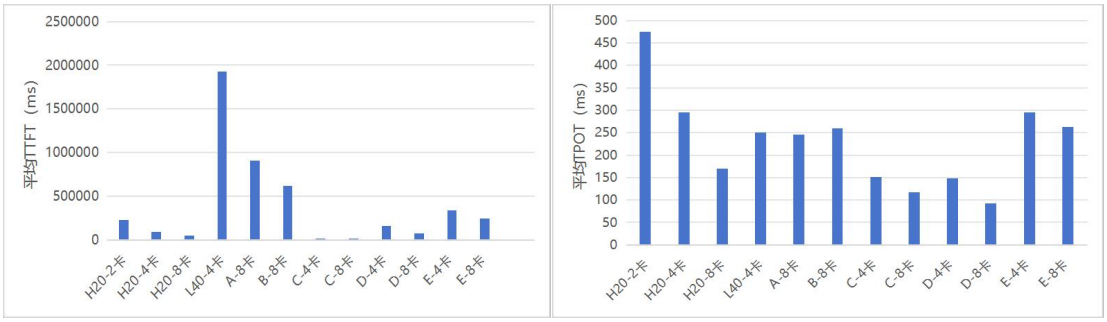

图 3-21　70B 512 并发的平均 TTFT（左）与 TPOT（右）

从测评结果来看，在小并发情况下，各种算力卡环境的 TTFT 与 TPOT 对比趋势基本一致，在大并发情况下，部分算力卡性能到达瓶颈、性能骤降，且对 TTFT 与 TPOT 的影响不完全一致，导致 TTFT 与 TPOT 的对比趋势出现不同。

以 32B 模型 32 并发的推理服务为例，在 SLA 设定平均 TTFT 小于 2s、TPOT 小于 100ms 的情况下，调度系统通过测评能够精准匹配可选资源列表，见表 3-3。

表 3-3 根据测评数据调度算力资源示例

TTFT 满足要求的资源列表	TPOT 满足要求的资源列表	TTFT 与 TPOT 均满足要求的资源列表
H20-8 卡、C-2 卡、C-4 卡、C-8 卡、D-8 卡	除了 A-4 卡、B-4 卡、B-8 卡，其余均满足	H20-8 卡、C-2 卡、C-4 卡、C-8 卡、D-8 卡

以上测评结果是在关闭 Prefix 缓存属性，或者每次测试前重启 vLLM 服务的情况下测得，不受 Prefix 缓存影响。

如果开启 Prefix 缓存属性，如图 3-22 所示是在 H20 及 E 卡上运行 32B 模型，进行 64 并发的首次推理与非首次推理 TTFT 对比，经过首次推理预热，非首次推理的 TTFT 将大幅下降，但是预热不会对 TPOT 性能产生如此跨数量级的剧烈影响。

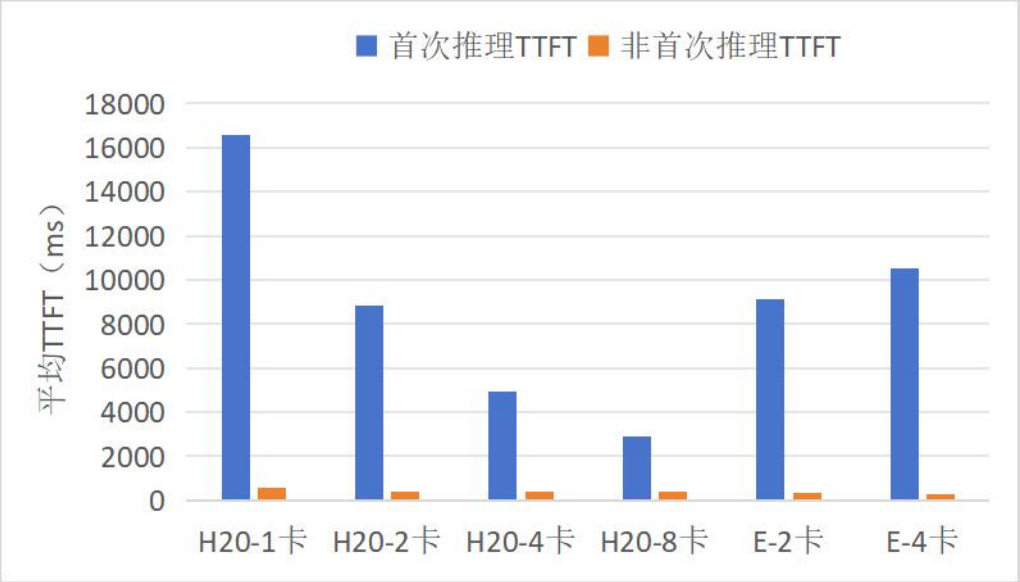

图 3-22　32B 64 并发的首次/非首次推理平均 TTFT（Prefix 缓存开启）

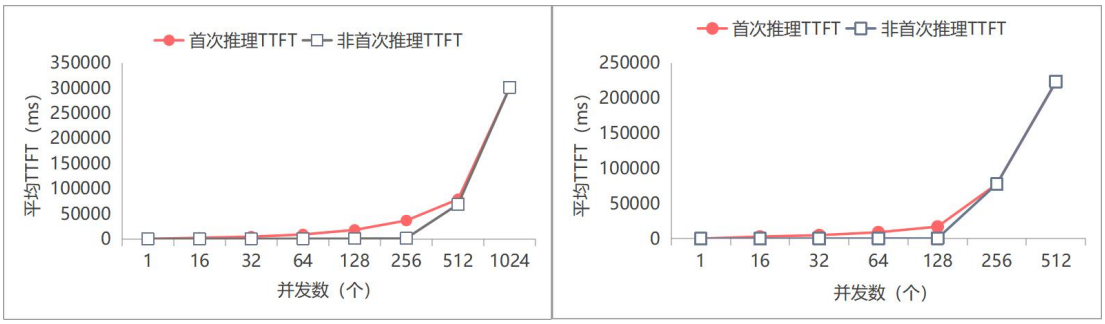

图 3-23　32B 2 卡运行 H20（左）与 E 卡（右）平均 TTFT（Prefix 缓存开启）

最后，我们针对 H20 与 E 卡开启 Prefix 属性，连续测试每个并发下的首次推理与非首次推理性能，每次更换并发数时重启 vLLM 服务，以保证首次推理性能不受 Prefix 缓存影响，测试结果如图 3-23 所示。可以看出，无论是首次推理还是非首次推理场景，平均 TTFT 在低并发情况下随并发缓慢增长，但是在高并发情况下，随着并发数增大超过算力承载能力，平均 TTFT 出现指数级激增，性能急剧下降；非首次推理由于高并发的 Prefix 缓存命中率降低，与首次推理的性能差异逐渐减小，并趋于统一。

四、典型场景与应用案例

4.1 入企 —— 交通规划报告

图 4-1 交通规划报告生成加速

本案例中，苏交科集团使用毕昇开源大模型应用开发平台与新华三一体机算力资源开展 DeepSeek-R1-Distill-Llama-671B 大模型本地推理服务，通过构建个性化智能体工作流，推理生成甘肃天水张家川县公路规划图文报告。

由于本地智算资源有限，当推理并发数超过阈值后，将发生资源抢占，影响推理速度。实测低并发情况下，生成报告耗时 42s，当推理并发数增大至 60，生成报告耗时增大至 72s，继续增大推理并发数至 100，生成报告耗时超过 300s。通过接入算力网调度与加速平台，实现推理业务云边协同负载，此时，本地计算压力被分流至全网可用算力节点，60 个并发的推理报告耗时缩短至约 45s，100 个并发数的推理报告耗时约 75s，推理速度和效率得到极大提升。

4.2 入企 —— 医疗问答推理

苏州某医疗研究所的权威医疗知识库，基于DeepSeek-R1-Distill-Llama-32B 大语言模型构建私有化医疗智能推理引擎，实现病理分析、用药推荐、诊疗路径推演等高阶医疗问答服务，为医生、科研人员及患者提供合规、可溯源的专业临床决策支持，显著提升医疗知识转化效率与诊疗准确性。

然而，在本地 L20 智能计算资源受限场景下，高并发推理请求易触发资源竞争瓶颈，导致服务时延激增、响应质量下降。通过引入算力网智能调度与加速能力，调用云端寒武纪 MLU370 算力进行推理加速，利用全局算力池化技术规避本地资源过载风险，实现推理效率与资源利用率的双重跃升。借助近源计算卸载与高速确定性网络，构建跨地域云边协同负载均衡能力：基于算力、时延、成本、碳排放等多目标优化动态分流策略，将本地推理压力自动分配至全网最优算力节点，降低端到端推理延迟 40%、提升并发吞吐量 3 倍、节省大模型部署成本 50%以上。

4.3 入园 —— 医疗诊断微调

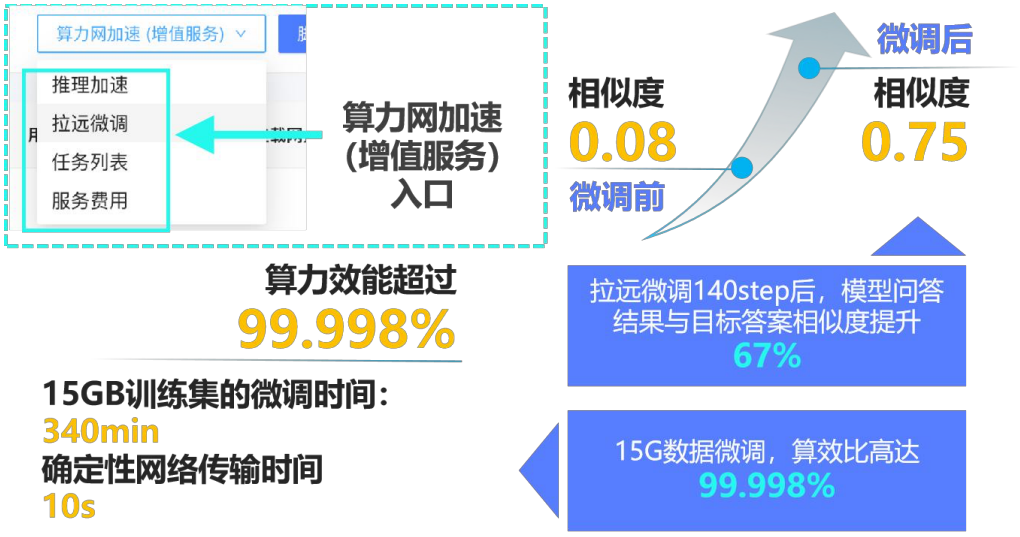

图 4-2 医疗诊断微调提高相似度

南京笑领科技有限公司的口腔医疗 SaaS 服务平台，使用 DeepSeek-R1-Distill-Llama-70B 医疗行业大模型提供推理问答服务。由于 DeepSeek 通用模型处理垂直领域问题效果欠佳，即使挂载专业知识库仍然不能满足医疗行业的专业推理需求，而笑领科技本地智铠 100 算力不支持企业模型微调升级，导致其 SaaS 服务平台推理问答业务发展受到严重影响。

依托算力网调度与加速平台，动态调度天数智芯宿州机房的天垓 150 算力，为笑领科技拓展拉远微调业务。实测微调前模型推理结果与测试集目标答案之间的平均相似度为 8%，使用 15GB 训练集微调 140 迭代次数后，平均相似度提升至 75%，有效降低大模型损失函数、提升推理准确度。

本案例中 15GB 训练集的微调时间为 5 小时 40 分钟，通过 11Gbps 确定性网络传输仅 10s，算力效能超过 99.998%。实测表明，确定性

网络的超高带宽可显著提升云边协同微调效率，通过压缩数据集传输时间，在相同服务等级协议（SLA）时限内最大化有效计算时长，从而允许选用低成本边缘算力执行微调任务，最终实现用户微调成本下降与效率提升的双重优化。

4.4 入校 —— 基因检测编辑

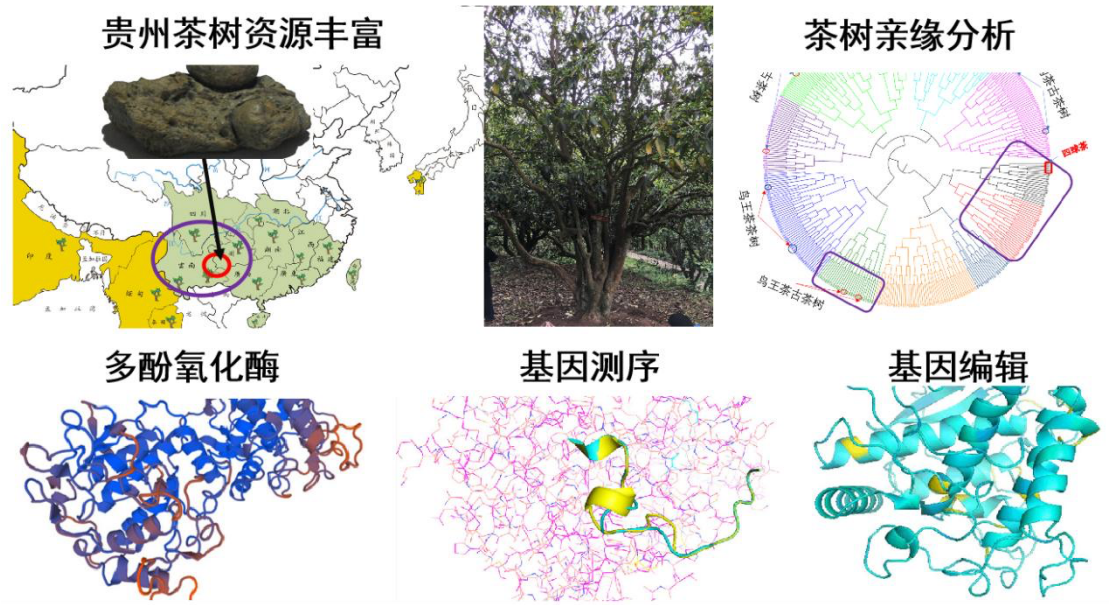

图 4-3 茶树多分氧化酶基因研究图示

贵州茶树资源丰富，曾发现世界唯一茶籽化石——四球茶籽化石，是茶树起源的核心区域和原产地中心之一，世界茶树各大聚类群都有贵州的茶树种质资源分布。开展贵州茶树的基因研究对全球茶叶品质加工、维持遗传种质多样性及杂交品种选育具有重要意义。

在本案例中，贵州师范大学（简称“贵师大”）师生选用多酚氧化酶（polyphenol oxidase, PPO）作为基因研究对象。PPO 可催化氧化儿茶素类物质形成茶黄素类（TFs）色素，对茶叶色、香、味等品质形成具有关键作用，是茶叶加工尤其是红茶品质形成的关键酶。

基因研究队列分析需处理百至上千样本的多组学数据，完成由10 个以上任务构成的分析链，基因组和蛋白结构分析及可视化需处理复杂序列比对，依赖高性能算力实现实时整合与交互分析，对时效性要求极高。贵师大本地算力不足，多组学研究亟需高性能智能算力，支撑茶树、荞麦等组学数据单次分析多达 10Tb 数据的行业大模型推理与结构比对研究。

通过接入算力网调度与加速平台，拉远拓展本地不支持的推理业务，赋能贵师大师生进行茶树多分氧化酶基因研究。使用 DeepSeek-R1-Distill-Qwen-32B 大模型推理进行多酚氧化酶基因测序，在此基础上编辑及比较基因序列。图 4-3 中基因测序部分为比较两个多酚氧化酶线段状结构图示，紫色部分为两个酶共有，黄色部分和蓝色部分分别为两个酶独有结构。通过研究茶树多酚氧化酶基因序列，贵师大师生精准识别出普安哈马四球茶、团龙贡茶古茶树等多样性重点或划片保护对象，根据亲缘远近关系选育出鸟王种茶树与鱼钩茶古茶树等优良杂交种，有效提升茶叶加工品质。

4.5 政务 —— 政务推理问答

江宁区数据局（政务办），负责贯彻落实省市区关于数据和政务服务管理工作的决策部署，以数据要素市场化配置改革为主线，统筹推进区域数字基础设施布局、数字经济、数字社会、数字政府规划和建设工作，牵头行政审批制度改革，优化政务服务等。

民生服务要求高，遇到特殊时间节点导致短时间激增的访问需求，

需要系统能实时精确的解答群众问题，避免等待时间长。但是本地算力不足，因此借助确定性网络无缝连接远方大模型算力中心，试点按需调配资源，降低整体成本。

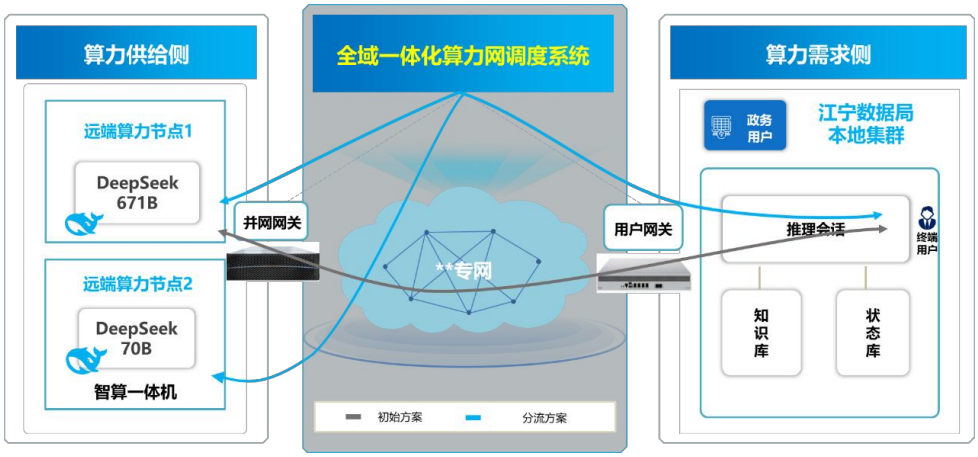

图 4-4 江宁数据局推理加速

如图 4-4 所示，基于算力网调度系统与政务通智能体，结合浪潮大模型一体机与远端算力资源，试点以本地部署的大模型作为计算核心枢纽，支持快速响应本地数据处理需求；当业务量激增，本地资源难以负荷时，可通过该平台灵活调用部署在云端的算力资源，保障智能体在高并发场景下的稳定运行，为推理用户提供及时、准确的服务。