CAICT 中国信通院

集智 专题报告

# 智能算力服务研究报告

## （2026 年）

中国信息通信研究院云计算与数字化研究所

2026年4月

# 版权声明

本报告版权属于中国信息通信研究院，并受法律保护。转载、摘编或利用其它方式使用本报告文字或者观点的，应注明“来源：中国信息通信研究院”。违反上述声明者，本院将追究其相关法律责任。

# 前　言

随着数字经济的蓬勃发展，算力已成为国家核心生产力和关键基础设施。当前的算力体系主要由通算、智算和超算三大类构成。以大模型为代表的人工智能技术爆发式演进，使智能算力需求呈指数级增长，成为全球科技竞争的焦点和算力体系建设的重中之重。但智算资源的跨域异构、网络传输瓶颈以及供需匹配痛点，日益制约着产业的发展。本报告正是立足于这一时代背景，重点针对智能算力展开深度探讨。

为破解算力资源协同难题，智能算力服务应运而生。**智能算力服务是基于互联网访问，汇聚 GPU、NPU 等异构计算资源，通过统一的服务接口，按需向用户提供可度量的计算、存储、网络等服务。**智能算力服务作为连接智能算力资源与上层行业应用的关键载体，正加速向**普惠化、智能化、精益化**方向演进，其发展水平直接关系到国家数字竞争力的全局。在此背景下，中国信息通信研究院云计算与数字化研究所基于在云服务、智能算力服务、算力互联网领域的长期研究积累，组织编写了**《智能算力服务研究报告（2026 年）》**，旨在系统梳理智能算力服务的内涵、体系架构、关键技术、典型场景、市场规模及发展趋势，为政策制定和产业实践提供参考。

本报告首次构建了**智能算力服务的三层体系架构**，包括**智能算力资源层服务、智能算力互联互通层服务以及智能算力应用层服务**。报告指出，在**算力标识、算网协同、资源池化及异构算力调度**等关键技术的支撑下，智能算力服务在**高强度计算与并行处理、结果交**

**付与体验类服务、垂直行业融合与边缘协同**等类型场景中不断深化；**全球智能算力服务市场保持稳健增长，我国市场增速尤为显著**，产业链上中下游新业态、新模式不断涌现。未来，智能算力服务将继续驱动产业呈现出**架构部署向“云边端高频协同”演变、服务模式向“任务式交付”升级、产业格局向“算力互联网”聚合、赋能路径向“生态价值”延展的四大核心发展趋势**。

期望本报告能够为政府部门、行业企业、研究机构及产业各界提供有益借鉴，共同推动智能算力服务成为像水电一样便捷可用的社会级基础性服务，赋能千行百业智能化转型升级。如有不足，恳请各界批评指正。后续我们将不断更新完善，如需探讨交流可发送邮件至 lipeishan@caict.ac.cn。

# 目 录

一、智能算力服务的发展背景 ..... 1
二、智能算力服务内涵 ..... 4
（一）智能算力服务的定义与突出特点 ..... 5
（二）智能算力服务的重要组成部分 ..... 7
三、智能算力服务体系架构 ..... 9
（一）智能算力体系架构 ..... 9
（二）智能算力服务体系架构 ..... 12
四、智能算力服务关键技术 ..... 16
（一）算力标识网关技术 ..... 16
（二）算网协同技术 ..... 17
（三）算力资源池化技术 ..... 19
（四）异构算力调度技术 ..... 20
五、智能算力服务典型场景 ..... 23
（一）高强度计算与并行处理场景 ..... 23
（二）结果交付与体验类服务场景 ..... 25
（三）垂直行业融合与边缘协同场景 ..... 26
六、智能算力服务市场与产业链分析 ..... 28
（一）市场规模分析 ..... 28
（二）产业链分析 ..... 30
七、智能算力服务发展趋势 ..... 35
（一）架构部署：从“中心化”向“云边端高频协同”演变 ..... 35
（二）服务模式：从“资源供给”向“任务式交付”升级 ..... 38
（三）产业格局：从“独立发展”向“算力互联网”聚合 ..... 40
（四）赋能路径：从“计算能力”向“生态价值”延展 ..... 41
附件：智能算力服务的典型场景 ..... 44
（一）长三角（上海）算力互联互通平台实践 ..... 44
（二）湖北算力互联互通平台运营实践 ..... 46
（三）深圳市算力互联网建设运营及赋能实践 ..... 48

（四）中国联通“CubeAI”算网模一体化服务平台实践 ..... 51
（五）成渝（四川）算力互联互通平台协同运营实践 .....53

CAICT 中国信通院

# 图 目 录

图 1 《算力互联互通行动计划》总体框架图 ..... 4
图 2 智能算力体系架构图 ..... 10
图 3 智能算力服务体系示意图 ..... 13
图 4 统一算力标识下的跨主体算力资源标准化互联与匹配 ..... 17
图 5 算网协同下的算力网络节点架构模型 ..... 18
图 6 资源解耦与可组合架构示意图 ..... 19
图 7 算网云调度系统 ..... 21
图 8 人工智能模型算力需求 ..... 35
图 9 “IDC+超节点”的协同范式 ..... 37
图 10 国家算力互联网服务平台架构图 ..... 45
图 11 湖北算力互联互通体系总体架构 ..... 47
图 12 深圳市智慧城市算力统筹调度平台架构图 ..... 50
图 13 全流程智算服务体系架构图 ..... 52
图 14 成渝（四川）算力互联互通体系架构图 ..... 54

# 表 目 录

表 1 全球主要国家算力政策动向总结 ..... 2
表 2 智能算力服务与云服务、IDC 服务存在三大区别和差异 ..... 7

## 一、智能算力服务的发展背景

在人工智能技术加速演进的当下，大模型的规模化落地更推动 AI 技术迈入规模化应用新阶段，算力也已从传统计算资源，逐步拓展演化到适配 AI 场景的智能化服务形态——智能算力服务。但与此同时，作为 AI 模型训练、推理及各类数字场景落地的核心支撑，智能算力服务的发展仍面临三大突出矛盾：**一是资源层，异构芯片适配协同难。**当前市场上芯片架构与软件生态较为封闭，企业若要在不同厂商的智算芯片间迁移模型，往往面临巨大的代码重构与适配成本。阻碍智能算力资源服务化协同与灵活调用。**二是互联互通层，算力供需错配问题。**随着大模型应用激增，推理需求快速增长，算力需求将从当前 70%以上来源于大模型训练所需的集中式算力，到未来 70%以上来源于推理所需的分布式算力[1]，发生深层次结构性转变，提出随时随地随需获取算力诉求。而我国算力供给在主体、区域等方面较为分散，分布式的算力需求高效匹配到分散的算力资源上难度较高。**三是应用层，算效有待提升。**既往以包机柜或租裸金属为主的粗放型资源租赁模式，已无法适应 AI 开发者对科研仿真、模型训练、视频渲染等计算任务更加精细化、结果导向的算力服务需求。面对三大突出矛盾，全球各国纷纷将智能算力服务竞争上升至国家战略高度，通过出台专项政策、加大资金投入，抢占智能算力服务发展的制高点。

**全球正加速推进“先建设算力，再推进标准规则互通，同步塑造**

[1] 数据来源：巴克莱预计，AI 推理计算需求将占通用人工智能总计算需求的 70%以上，甚至可以超过训练计算需求，达到后者的 4.5 倍。 IDC 数据，到 2026 年，云端部署的算力中推理占比将从 2022 年的 58.5%提升至 62.2%，而训练算力占比则下降至 37.8%。此外，IDC 进一步预测，到 2025 年，中国智能算力内部结构中，训练算力占比将下滑至 27.4%，推理算力占比将上升到 72.6%。

**全球算力服务大棋局**”。主要国家纷纷出台政策，其核心方向高度一致地聚焦于推动算力互联、优化资源调度，以构建国家层面的算力服务优势。美国凭借《赢得 AI 竞赛：AI 行动计划》和“星际之门”等政策和项目，计划投入高达 5000 亿美元，构建以自身为核心的全球逻辑算力互联网，强化其技术领导地位；欧盟为摆脱“技术消费者”的被动角色（当前自有算力仅占全球 3%-5%），推出“人工智能大陆行动计划”并计划立法，投入 2000 亿欧元，目标在 2032 年前将算力提升三倍，建立覆盖多国的泛欧主权 AI 算力逻辑网络；英国通过“人工智能机遇行动计划”投入 140 亿英镑，力求在 2032 年前将 AI 算力扩大 20 倍以减少对美依赖；法国则依据其《国家 AI 战略》，凭借能源优势投入 1090 亿欧元，以将自己打造为“欧洲算力枢纽”和全球 AI 基础设施托管中心；中东国家如沙特和阿联酋，则利用主权基金优势，通过沙特新未来城（NEOM）等项目积极承接美国技术转移，致力于建设区域算力枢纽和数字新兴场景。

表 1 全球主要国家算力政策动向总结

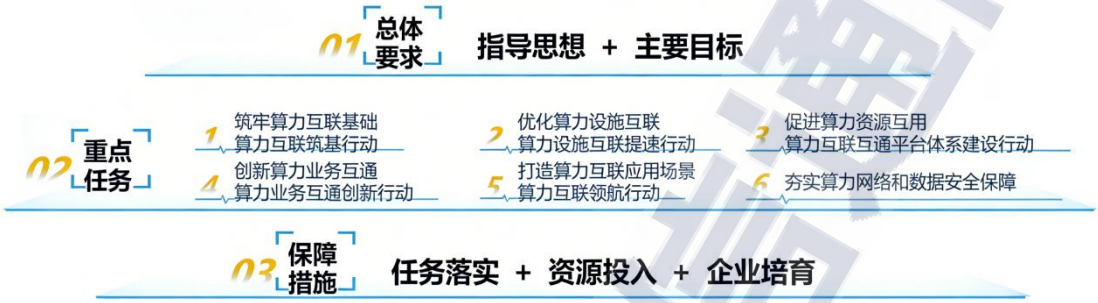
| 国家/地区 | 核心战略 | 政策文件/项目 | 投资规模 | 目标 |
| --- | --- | --- | --- | --- |
| 美国 | 构建美系全球算力互联网 | 《赢得 AI 竞赛：AI 行动计划》、"星际之门"以及"OpenAIforCountries"项目(2025) | 5000 亿美元 | 实现全美算力互联，以 10 国为起点，部署全球算力互联网 |
| 欧盟 | 建立泛欧主权 AI 算力。当前欧洲仅占全球算力的 3%—5%，需通过大规模投资将比例提升，以摆脱“技术消费者”的被动地位。 | “人工智能大陆行动计划”（AIContinentActionPlan)(2025)《云和 AI 发展法案》（计划中） | 2000 亿欧元 | 2032 年算力翻 3 倍，覆盖 17 国 13 个 AI 工厂 |

| | | | | |
|---|---|---|---|---|
| 英国 | 摆脱对美云依赖，追赶美中算力、AI 水平 | “人工智能机遇行动计划”（2025）《计算路线图》(2025) | 140 亿英镑；20 亿英镑 | 2032 年前 AI 算力扩 20 倍，新建爱丁堡超算中心 |
| 法国 | 将法国打造为全球 AI 基础设施托管中心，摆脱对中美技术的依赖，并争夺全球 AI 领导权。 | 法国《国家 AI 战略》第三阶段 | 1090 亿欧元 | 利用其独特的能源优势，将法国打造成为“欧洲算力枢纽” |
| 中东国家 | 打造区域算力枢纽 | NEOM（沙特）、“星际之门阿联酋”（阿联酋） | 未公开（主权基金主导） | 承接美企技术转移，建设数字孪生城市等新兴场景 |

来源：中国信通院根据公开信息整理

**中国正系统推进算力“先互联，再成网，同步构建全国统一大市场”**。国务院在政府工作报告中强调“加快形成全国一体化算力体系”，发改委、工信部等五部门联合印发《关于深入实施“东数西算”工程加快构建全国一体化算力网的实施意见》明确提出“加强多元算力互联互通和统一服务”。2025 年，工信部印发《算力互联互通行动计划》（以下简称《行动计划》）标志着我国算力互联互通进入标准化、体系化推进新阶段。《行动计划》按照“先互联、再成网、同步建市场”的总体思路，开展“推动算力互联互通，构建算力互联网，同步打造统一算力服务的大市场”系列工作，指导产业界协同推动算力服务向更加普惠化、智能化、精益化发展。《行动计划》提出**到 2026 年**，建立较为完备的算力互联互通标准、标识和规则体系。设施互联方面，推广新型高性能传输协议，提升算力节点间网络互联互通水平；资源互用方面，建成国家、区域、行业算力互联互通平台，统一汇聚公共算力标识，实现全国头部算力企业的公共算力资源互联；业务互通方

面，推动算、存、网多种业务互通，实现跨主体、跨架构、跨地域算力供需调度；应用场景方面，开展算力互联网试验网试点，赋能产业普惠用算。**到 2028 年**，基本实现全国公共算力标准化互联，逐步形成具备智能感知、实时发现、随需获取的算力互联网。




来源：中国信息通信研究院

图 1《算力互联互通行动计划》总体框架图

总体来说，从全球各国的战略博弈到我国一体化算力体系的构建，本质上都是对“高效用算、互联互通”核心需求的回应，而产业关注点从拥有算力到用好算力，更标志着智能算力服务进入**普惠化、智能化、精益化**的全新发展阶段，为人工智能规模化赋能产业、激活数字经济新动能奠定坚实基础。

## 二、智能算力服务内涵

在梳理全球主要国家推进智能算力服务的战略动向以及我国政策布局基础上，理解智能算力服务内涵，关键在于从概念边界、服务特征和演进逻辑等维度，对其内涵进行系统界定；并在此基础上，界定智能算力服务的重要组成部分，厘清其与云服务、互联网数据中心（Internet Data Center，以下简称为 IDC）服务的差异，明确其从物理资源聚集向标准化服务交付跃升的本质特征，从而为全文的深入探讨

提供坚实的基础认知框架。

## （一）智能算力服务的定义与突出特点

### 1.智能算力服务定义

**智能算力服务是基于互联网访问，汇聚 GPU、NPU 等异构计算资源，通过统一的服务接口，按需向用户提供可度量的计算、存储、网络等服务。**作为一种新型的服务形态，智能算力服务旨在破解当前产业面临的跨地域物理隔离、跨架构生态割裂与供需匹配低效等核心痛点。它通过屏蔽底层基础设施及异构硬件的复杂性，打破传统物理边界的限制，将各地分散的算力资源转化为可全域流动、随需取用的标准化服务能力。

### 2.智能算力服务突出特点

**智能算力服务、云服务与 IDC 服务，代表了算力产业在不同发展阶段的服务形态。**理解智能算力服务发展沿革与三者差异，需明确其内在的演进逻辑与业务边界。

根据《电信业务分类目录（2015 年版）》定义，IDC 业务经营者应提供机房和相应的配套设施，并提供安全保障措施。智能算力服务不包含 IDC 业务内容，指的是智能算力资源服务、智能算力互联互通服务、智能算力应用服务。

尽管智能算力服务涵盖并兼容了云服务的基础能力，但在面向新兴的、复杂的计算任务时，它与云服务、IDC 服务在核心逻辑上表现出三大显著区别。

（1）交付形态不同

IDC 服务受限于物理设施，交付核心在于机柜与机房空间，用户需自行关注底层硬件的运维与管理；云服务虽然实现了资源的虚拟化，将交付单元升级为标准化的云主机与云存储，但其本质仍是资源的供给；相比于云计算以“资源租赁”为主的模式，智能算力服务实现了向“任务式交付”的重心跃迁。它不再局限于单一的资源交付，而是可直接以“算力词元（Token）”、“训练任务”、“渲染作业”等计算任务为交付形式。通过算力互联网的智能调度，服务能够穿透资源层直接响应业务目标，实现了价值的精准输出。

（2）技术重心不同

IDC 服务的核心在于保障物理基础设施的稳定可靠，技术重心主要落在土木建设、电力供应、制冷系统等工程层面；云服务的技术核心在于资源池化与虚拟化，重点解决同构资源的弹性管理与按需分配；而智能算力服务面临着更为复杂的挑战，其核心技术聚焦于异构计算、智能调度与高速互联。依托算力互联网的统一标识符与调度操作系统，它旨在高效整合并输出 CPU、GPU、NPU 等跨地域、跨架构的多元算力，解决异构资源协同难的问题。

（3）计费逻辑不同

IDC 服务延续了资产租赁模式，主要按机柜数量、空间占用和带宽规模计费；云服务虽然进化为按需付费，依据资源规格和使用时长进行结算，但本质上仍体现了资源的分时租赁；智能算力服务的商业模式则更加灵活多元，可以采用按计算任务复杂度、算力实际产出或

精确的算力使用量进行计费。这种模式将服务成本与用户的业务成果更紧密地绑定，体现了从购买资源向购买效果转变的服务理念。

表 2 智能算力服务与云服务、IDC 服务存在三大区别和差异

| 特性 | 智能算力服务 | 云服务 | IDC 服务 |
| --- | --- | --- | --- |
| 交付形式 | 算力词元（机时、卡时、训练任务、渲染作业等 | 云主机、云存储、数据库等 | 物理机柜、机房等 |
| 技术重心 | 异构计算、智能调度、高速互联 | 资源池化、虚拟化 | 土木、电力、制冷 |
| 计费逻辑 | 按计算任务复杂度或算力使用量 | 按资源规格和使用时长 | 按机柜/空间/带宽 |

来源：中国信息通信研究院

## （二）智能算力服务的重要组成部分

智能算力服务强调无论底层物理设施产权归属，所构建和提供的、可被调度与交易的虚拟化资源与服务，以及基于虚拟化的算力资源，进行调度与交易，并且实现最终的应用服务。

基于当下发展重点，词元服务、算力互联网服务、云服务是智能算力服务的比较典型的组成部分，三类服务形态相互衔接、互为补充，共同构建了从通用资源租赁、跨域算力调度到任务式能力交付的完整服务生态。

### 1.词元服务

词元服务是智能算力服务在任务化交付阶段形成的新型服务形态。其通过对分散的模型调用能力和底层算力资源进行统一封装，并以词元作为标准化计量单位对外输出，实现跨地域、跨主体、跨架构资源的聚合供给与灵活调用。相较于云服务侧重资源交付、算力互联网服务侧重资源调度，词元服务更强调交付结果的任务式服务。

### 2.算力互联网服务

算力互联网服务是面向分布式推理、大规模协同训练等复杂场景的高阶形态。针对智能算力资源分布不均与需求爆发之间的矛盾，算力互联网服务在互联网体系架构之上，引入了统一算力标识与算网云调度操作系统。它打破了物理边界，将各地分散的异构算力编织成一张逻辑网络，具备智能感知、实时发现与随需获取能力。通过算力互联网，用户可以像使用水电一样，灵活调用跨地域、跨架构的智能算力，实现“一点接入、全算贯通”。

### 3.云服务

云服务是智能算力服务的基础交付形态。主要面向标准化、同构的虚拟化资源需求，通过资源池化和虚拟化技术，提供弹性、按需的基础设施即服务（Infrastructure as a Service，以下简称 IaaS）。云服务解决了基础算力资源的规模化供给问题，是当前 AI 企业获取算力的主要方式。

# 三、智能算力服务体系架构

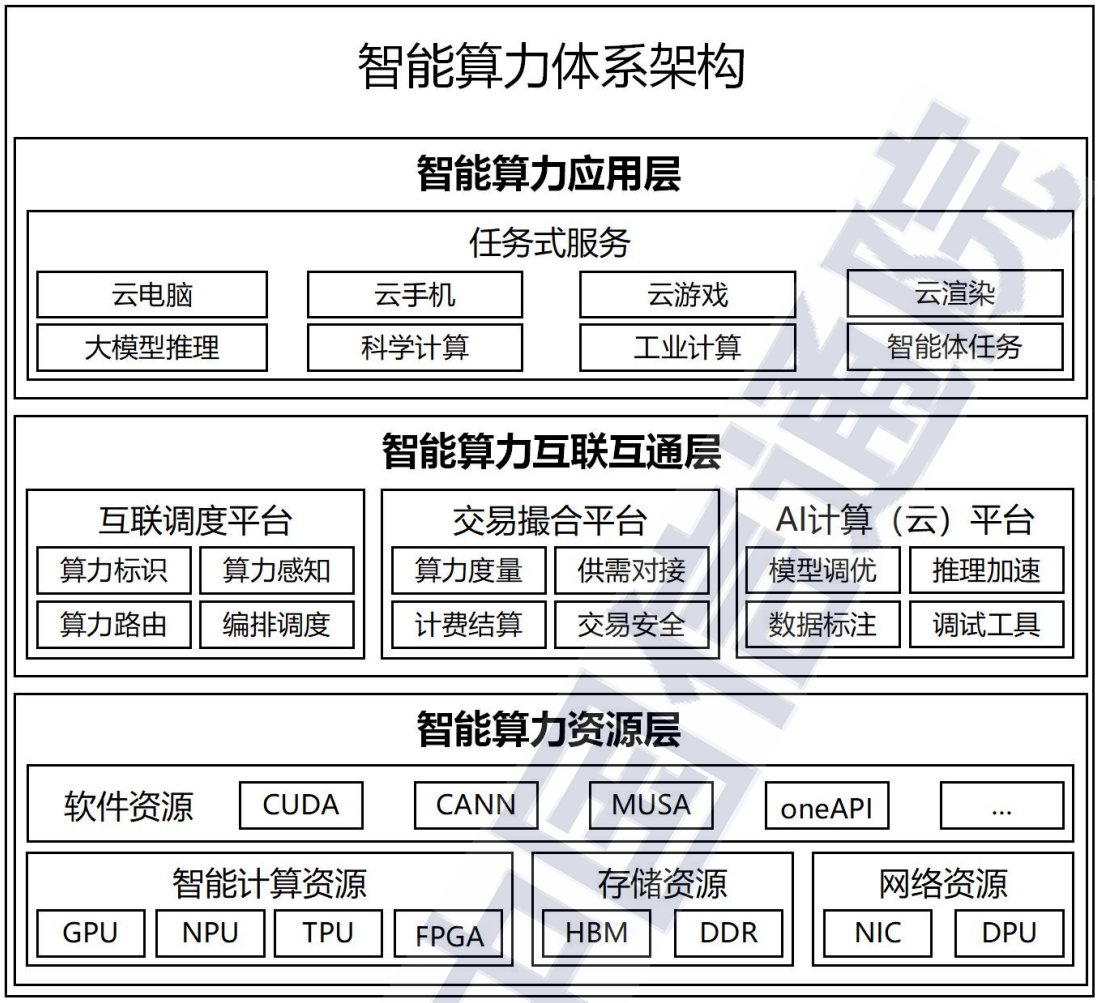
在明确智能算力服务内涵的基础上，进一步理解智能算力服务的产业形态，还需要从“体系架构”与“服务映射”两个维度展开分析：一方面，明确智能算力资源层、智能算力互联互通层、智能算力应用层的功能定位及其相互关系；另一方面，进一步将上述架构分层映射为智能算力资源服务、智能算力互联互通服务、智能算力应用服务三类服务形态，厘清不同服务的供给边界、运行逻辑与交付方式，从而为后文分析智能算力服务关键技术、典型场景及产业趋势奠定结构化认知基础。

## （一）智能算力体系架构

面对人工智能大模型训练与推理带来的指数级算力需求增长，以 CPU 为核心、各层紧耦合的计算体系已无法适应异构资源的高效调度与大规模并行处理需求。为此，我们提出了一套面向“智算服务化”的全新智能算力体系架构。

架构的设计逻辑遵循“服务交付颗粒度”与“产业价值链解耦”的科学原则，旨在解决“算力资源的异构复杂性”与“用户使用的便捷性”之间的矛盾。我们将体系划分为三个逻辑层级：**智能算力资源层、智能算力互联互通层与智能算力应用层**。

智能算力资源层提供标准化的异构算力供给服务，智能算力互联互通层通过“卡时”计量实现高效的调度与交易服务，智能算力应用层则最终面向用户提供“任务式”的价值交付。三层架构协同运作，将底层的运算能力转化为可流动的**智能算力服务**。




来源：中国信息通信研究院

图 2 智能算力体系架构图

## 1.智能算力资源层

**智能算力资源层**提供以 GPU 等芯片、存储、网络等软硬件资源为核心的智能算力资源集合，支撑大模型等训推的集群能力。基模竞争（Scaling Law）依然生效，万卡集群在超万亿参数、多模态大模型训练仍发挥不可替代作用。

**智能算力资源层是智能算力服务的供给单元。**面对底层 CPU、GPU、NPU、FPGA 等高度异构的硬件设施，不同于 IDC 简单的物理设施托管，智能算力资源层致力于将异构资源进行池化与抽象化，形

成统一的智能算力资源池。在大模型时代，本层重点强化了集群服务能力，能够提供支持千卡、万卡级并行的训练环境，满足 AI 大模型对算力规模和稳定性的苛刻要求。

**2.**智能算力互联互通层

**智能算力互联互通层**基于智能算力资源提供 AI 计算、资源互联调度、供需交易撮合等能力的互联互通服务，决定人工智能随时随地随需使用算力服务。

**智能算力互联互通层是智能算力服务的运营载体。**基于智能算力资源层提供的基础能力，通过 1+M+N 国家、区域、行业算力互联互通平台体系，解决算力资源“找、调、用”的痛点，实现从资源供给向服务赋能的转变。作为调度中枢，通过统一的算力标识符、算网云调度操作系统和高性能传输协议，它将分散、异构的资源在逻辑上整合为统一的“算力互联网”。实现跨主体、跨架构算力资源的标准化互联与感知，让算力交易采用“卡时（GPU-hour）”等标准化单位进行计量，促进了算力资源的高效供需匹配与流动。

**3.**智能算力应用层

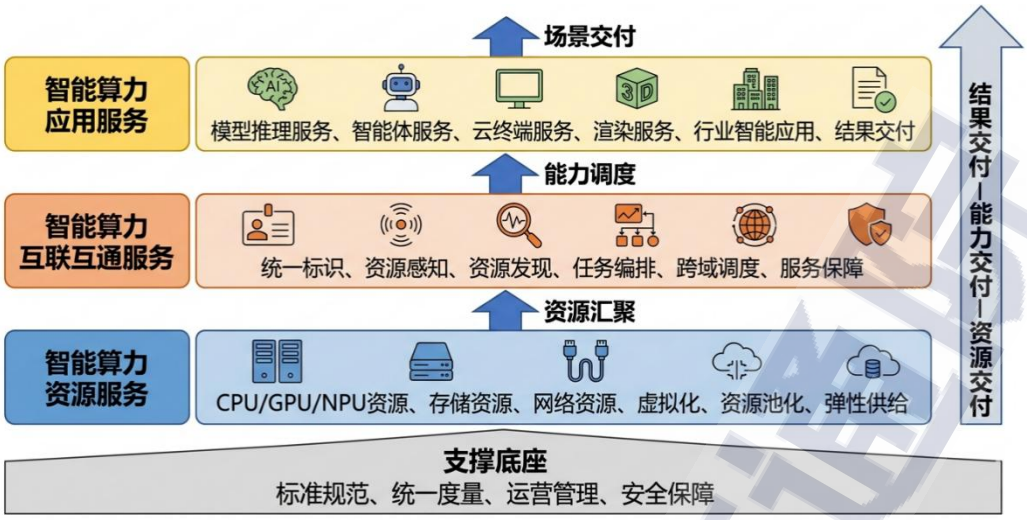
**智能算力应用层**面向人工智能、科学计算等场景，形成云电脑、云手机、大模型推理、视频渲染、智能体等企业级和消费级应用，用户使用智能体、视频渲染、云游戏等应用时可以实现智能算力“漫游”。

**智能算力应用层是智能算力服务的交付窗口。**智能算力服务应用层直接面向行业场景，将底层的智能算力资源与中间层的互联互通智能调度能力，封装成可快速部署、便捷调用的业务解决方案。它改变

了用户必须自行搭建环境、配置参数的模式，让用户能够聚焦于业务本身，推动服务模式由获取底层资源向面向任务成效的价值交付转变，让智能算力真正成为赋能千行百业的核心生产力。在具体的服务落地中，智能算力服务应用层依据不同的业务需求衍生出了多元化的服务形态，包括：追求极致算力的高强度计算与并行处理、注重即时反馈的结果交付与体验服务，以及深度嵌入生产流程的垂直行业融合应用，构成了智能算力服务丰富多样的典型应用场景。

## （二）智能算力服务体系架构

基于智能算力体系架构，智能算力服务体系是对底层资源能力、跨域调度能力和上层应用能力的服务化映射。智能算力服务体系并不是将资源、网络和应用进行简单叠加，而是在标准规范、统一度量、运营管理、安全保障等基础能力支撑下，形成**智能算力资源服务**、**智能算力互联互通服务**、**智能算力应用服务**三大服务层级。其中，**智能算力资源服务**负责将异构算力资源转化为可供给的基础能力，**智能算力互联互通服务**负责打通跨主体、跨地域、跨架构的互联互通与调度通道，**智能算力应用服务**则面向具体场景完成服务封装与任务交付，三者相互衔接、协同运转，共同构成智能算力服务的完整体系。




来源：中国信息通信研究院

图 3 智能算力服务体系示意图

## 1.智能算力资源服务

智能算力资源服务是整个体系的基础供给单元，主要面向 CPU、GPU、NPU 等计算资源，以及存储资源、网络资源等底层能力，依托虚拟化、资源池化等技术，形成可统一管理、弹性调配、按需供给的资源服务形态。其核心在于将原本分散、异构、难以直接使用的底层设施进行抽象和封装，转化为具备标准接口和统一度量方式的基础资源能力，为模型训练、模型推理、科学计算、视频渲染等任务提供稳定的算力底座。

智能算力资源服务并不局限于单一芯片或单一节点的供给，而是强调对多类型资源的统筹组织与集约输出。通过对计算、存储、网络等要素进行协同配置，可以构建面向千卡级、万卡级集群的弹性服务环境，满足不同场景下对算力规模、并发能力、可靠性和可扩展性的差异化要求。也就是说，智能算力资源服务解决的是“有没有资源、

资源能不能形成规模化供给”的问题，是整个智能算力服务体系得以建立的前提。

### 2.智能算力互联互通服务

智能算力互联互通服务是智能算力服务体系的中枢运营载体，承担着连接资源供给侧与应用需求侧的重要作用。其核心并不只是把资源“连起来”，而是围绕统一标识、资源感知、资源发现、任务编排、跨域调度、服务保障等关键环节，形成一套面向全网资源的组织、协同与运营能力。通过这一层服务，原本分散在不同地区、不同主体、不同技术架构下的算力资源，得以被统一识别、动态感知、实时发现和有序调用，从而由“资源分布”转变为“能力联动”。

从服务属性看，智能算力互联互通服务本质上是将算力资源从“可供给”进一步推进到“可调度、可流通、可保障”的阶段。一方面，它依托统一算力标识和调度系统，提升用户“找算力、调算力、用算力”的便捷性和精准性；另一方面，它通过跨域编排和标准化度量，推动算力资源在更大范围内实现高效匹配与有序流动。可以说，这一层服务解决的是“资源能不能高效协同、能不能跨域使用、能不能稳定交付”的问题，是智能算力由资源能力走向服务能力的关键环节。

### 3.智能算力应用服务

智能算力应用服务是智能算力服务体系面向用户和行业场景的最终交付窗口。其主要作用是将底层资源能力与中间层调度能力进一步封装为可直接使用的应用形态和服务方案，包括模型推理服务、智能体服务、云终端服务、渲染服务以及面向各行业的智能化应用等。

相较于资源租赁或能力调用，应用服务更加强调围绕具体业务需求完成任务交付和结果输出，使用户无需理解底层算力结构、网络路径和调度机制，即可直接获得可使用、可消费、可评价的服务成果。

从演进趋势看，智能算力应用服务代表了智能算力服务向高阶形态发展的方向。它推动服务模式从“提供资源”向“提供能力”、再向“提供结果”演进，体现出智能算力服务对底层复杂性的持续屏蔽和对业务价值的持续贴近。例如，一些 AI 应用服务商采用“算力豆”等计费方式，用户用资金购买“算力豆”，完成一个视频渲染任务后，扣除相应所需的“算力豆”。在这一层中，用户关注的重点已不再是具体调用了哪种芯片、占用了多少机时，而是一次模型推理是否高效完成、一次智能体任务是否顺利执行、一项渲染作业是否如期交付。推动对算力资源利用效率和服务质量提升，实现“随时、随地、随需地接入算力资源”。由此，智能算力真正从后台基础能力转化为前台可感知、可使用、可创造价值的生产服务。

# 四、智能算力服务关键技术

智能算力服务的落地，依赖于计算技术与网络技术的深度融合与系统性创新。目前智能算力服务的关键技术正呈现出从“单点突破”向“全栈协同”演进的总体趋势。为支撑算力资源跨主体、跨地域的高效互联与精准配置，其技术体系主要围绕四大核心环节构建。

**首先是“标”，**即统一认知，通过算力标识技术为全网资源建立“身份证”，实现算力的标准化感知与注册。**其次是“联”，**即高速传输，解决的是高效连通问题，侧重于底层数据的高速传输与低时延通道保障，主要依托算网协同技术来打破网络瓶颈，为算力构建低时延、大带宽的传送动脉。**再次是“池”，**即资源聚合，利用算力资源池化技术打破物理边界，将异构算力抽象为统一、弹性的逻辑资源池。**最后是“调”，**即智能配置，侧重于上层计算任务与各大品牌不同架构底层算力芯片的精准匹配与高效分发，依靠异构算力调度技术构建全网操作系统，实现任务与资源的最优匹配与编排。

这四大技术群相互支撑，共同构成了智能算力服务的坚实底座，确保算力资源能被高效、精准地输送给各类应用。以下将对各项技术展开详细论述。

## （一）算力标识网关技术

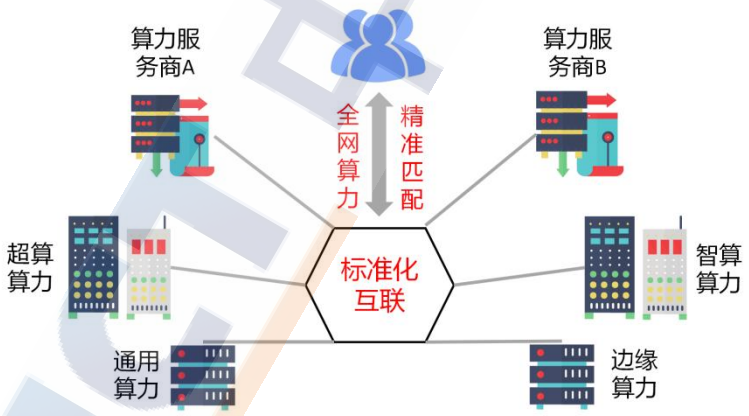
算力标识赋予了算力资源全局唯一的身份和多维属性描述，解决了用户“找算力”的难题。它使得网络不仅能识别 IP，还能识别算力类型、规格和实时负载。通过规范标识编码和升级算力网关，实现了各主体算力资源的标准化互联，保障了算力间可查询、可对话、可调用。

### 1.算力标识

算力互联网引入类似互联网 URL 的定位机制，即算力互联网标识（C-URL），用于在应用层基于统一标识码，定位算力资源和路径（算力路径）。当前研究重点在于规范算力标识编码方式，将资源情况、位置信息、规格信息、路径情况等自上而下的信息进行编码，从而提升应用层匹配合适的算力资源和路径的效率。

### 2.算力网关

算力网关作为连接用户请求和算力资源的入口，负责协议转换和路由分发。通过升级算力网关，并配合统一算力标识体系，可以将各算力提供主体的公共算力资源进行标准化互联，实现算力间可查询、可对话、可调用，使用户能便捷地使用多种算力资源。



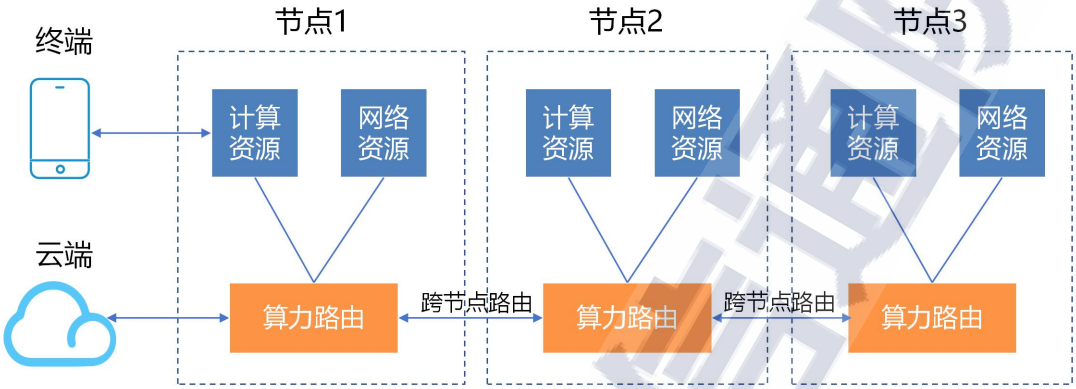

来源：中国信息通信研究院

图 4 统一算力标识下的跨主体算力资源标准化互联与匹配

## （二）算网协同技术

算网协同打破了网络与计算的孤立，将算力状态引入网络决策。它通过增强协议架构和高速互联能力，保障了分布式智能训练任务中

数据和算力应用的高速、低时延流动。同时，它利用网络的可编程性动态优化传输路径，确保算力能以最优链路被调用，是实现“算力随需取用”的基础。




来源：中国信息通信研究院

图 5 算网协同下的算力网络节点架构模型

### 1.远程直接内存访问（RDMA）

RDMA 允许一台计算机的内存直接访问另一台计算机的内存，无需操作系统内核参与，释放了 CPU 算力并大幅降低通信时延。目前以 RoCEv2（RDMA over Converged Ethernet）协议为主流，它使得 RDMA 能够运行在以太网上。在高性能计算（HPC）和大规模 AI 集群中，RDMA 已成为 GPU 之间高速互联的标准配置，显著解决了分布式训练中的通信瓶颈。

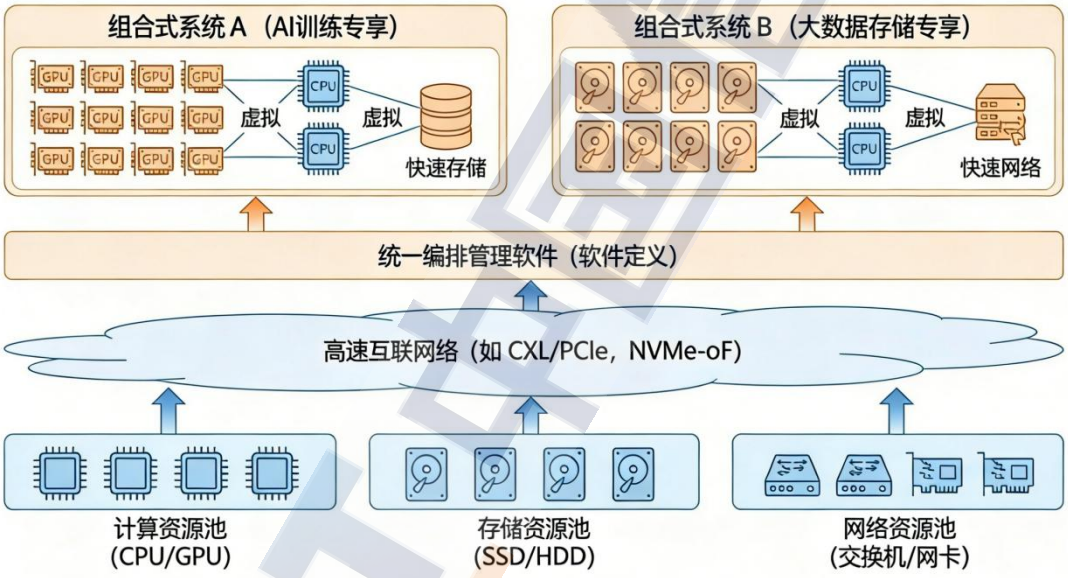
### 2.基于 IPv6 的段路由（SRv6）

SRv6 基于源路由理念，将网络路径和指令编码在 IPv6 报头中。它不仅能指定数据包的传输路径，还能携带算力信息和服务指令，实现网络的可编程性。SRv6 在算力网络骨干网调度中应用，实现基于骨干网直联点和新型互联网交换中心（IXP）的算力交换和寻址，是

强化网络设施间互联能力、提升算力调度速度的关键技术。

## （三）算力资源池化技术

资源池化将物理分散的计算、存储和网络资源进行抽象汇聚，形成逻辑统一、可灵活调度的资源大池。它通过容器/虚拟化技术实现了资源的高效共享和多租户隔离，显著提高了 GPU 等昂贵异构算力的利用率。结合资源解耦和 DPU，它加速了数据处理，为智能算力提供弹性、高可用的硬件基础。




来源：中国信息通信研究院

图 6 资源解耦与可组合架构示意图

### 1.虚拟化/容器

虚拟化通过虚拟机监视器（Hypervisor）或容器引擎，在物理硬件和操作系统之间引入抽象层，将物理资源切分为多个独立的逻辑单元，实现应用与硬件解耦。容器（Docker）、K8s（Kubernetes）等容器技术因其轻量级和敏捷性，已成为云原生算力的主流交付方式。当

前重点在于“神龙架构”类技术的演进，即通过软硬协同（智能网卡卸载虚拟化开销）来实现接近裸金属性能的虚拟化，解决虚拟化带来的性能损耗问题。

2.资源解耦

资源解耦打破了服务器以 CPU 为中心的封闭架构，将计算、存储、内存等硬件部件独立封装，通过高速总线互联，根据业务需求动态组合硬件资源。高速互联协议（Compute Express Link）技术是目前最大的突破点，它提供了内存一致性的高速互联标准。业界正在从“机架级解耦”向“数据中心级解耦”演进，旨在构建完全可组合的基础设施（Composable Infrastructure），解决大模型训练中显存墙和内存墙的问题。

## （四）异构算力调度技术

异构算力调度技术通过构建先进的算网云调度系统，实现了对不同架构、跨云资源的统一编排和高效分配。它通过智能算法和统一适配层，实现了任务与最优硬件的精准匹配，解决了“调算力”和“用算力”的难题。这保证了在复杂、多样的算力环境中，智能应用能够跨架构平滑迁移，并最大化集群整体效能。

来源：中国信息通信研究院

图 7 算网云调度系统

## 1.统一资源管理框架

统一资源管理框架主要体现为构建算网云调度系统，类似于云计算的 K8s 集群编排调度系统，通过分层架构实现从资源管理到应用层的自动化编排。当前进展是攻关实现对异构算力、异构网络和跨云资源的统一调度，其调度能力直接决定了算力任务调度的规模和效率，是解决“调算力”难题的核心。

## 2.任务画像与匹配算法

通过在应用层和底层硬件之间构建统一适配层，实现应用跨架构、跨主体的统一开发部署。当前进展是先通过发展兼容开放计算语言（Open Computing Language，简称 OpenCL）、CUDA 等主流生态的 OGPU（One-GPU）计算框架和 ODPU（One-DPU）开发框架解决 GPU 生态兼容的难点，然后通过统一接口简化应用部署，最终提高“用

算力”的计算部署效率。

CAICT 中国信通院

# 五、智能算力服务典型场景

在服务体系和关键技术逐步成型后，智能算力服务的价值最终要通过具体场景来体现。不同场景对算力规模、时延、交付方式和部署架构的要求存在显著差异，也正是在这些差异化需求牵引下，智能算力服务不断从通用供给走向精益化交付。本章从计算性能、交付模式和架构融合三个互补维度出发，系统展示智能算力服务如何在高强度计算、结果交付与体验服务、垂直行业融合与边缘协同等场景中释放价值。

## （一）高强度计算与并行处理场景

此类场景聚焦智能算力服务的极限性能，以大规模集群和并行计算为特征，是当前推动算力基础设施升级的主要驱动力，强调对算力规模、并发性与吞吐量的“压强”式需求。

### 1.AI 大模型训练与推理

AI 大模型场景涵盖了从千亿参数模型的预训练到高并发推理的全生命周期。在国际上，OpenAI 依托微软 Azure 构建的超大规模智算集群，通过数万张 GPU 的并行计算完成了 GPT-4 的训练；在国内，百度“文心一言”、华为“盘古”等大模型均依托智算中心实现了月级到周级的训练加速。

智能算力服务在此场景中通过“算网模一体化”大幅降低了企业门槛。例如，中国联通“CubeAI”平台（详见附录四），通过异构算力与网络模型的深度融合，有效解决了训练中网络丢包导致的断点续训

问题。服务商通过云端弹性伸缩，能有效应对推理流量的波峰波谷，如 ChatGPT 在流量爆发期通过动态扩容承载了亿级用户访问。

### 2.科学计算与仿真

科学计算（AI for Science）正利用混合算力架构解决气象、医药、流体力学等领域的难题。典型案例如 DeepMind 团队利用 AlphaFold 2 成功预测了自然界几乎所有的蛋白质结构，将原本需要数年的实验工作压缩至数天。在国内，国家气象局利用智算算力将气象数值预报的精度提升至公里级。

科学计算与仿真场景极度依赖高精度双精度浮点运算（FP64）。上海超算中心及长三角算力互联平台（详见附录一）已通过算力调度广泛服务于高校与科研机构，支撑了新药筛选与天体物理模拟等任务，避免了科研单位自建高昂算力设施的内存墙瓶颈。

### 3.高清视觉渲染

影视特效与工业设计属于典型的“潮汐式”图形密集型任务。电影《阿凡达：水之道》的制作调用了 AWS 云端海量的算力实例进行光线追踪渲染。国内电影《流浪地球 2》在制作过程中，同样利用云端渲染农场完成了数以万计的高难度视效镜头。

通过智能算力服务，原本需要本地工作站数周完成的任务被拆解至云端成千上万个节点并行处理，交付周期被压缩至小时级。深圳市算力互联网（详见附录三）整合了区域内的闲置 GPU 资源，为大湾区的动漫与设计企业提供了低成本、高效率的离线渲染服务，显著降低了数字内容创作的成本。

## （二）结果交付与体验类服务场景

这类场景代表了智能算力服务的创新形态，用户关注点从购买和管理底层计算资源彻底转向购买最终的结果或体验。服务商通过 API 或流化技术，完全屏蔽了底层技术的复杂性。

### 1.云游戏

云游戏将 3A 大作的运行与渲染完全迁移至云端，打破了终端硬件限制。国际上，微软 Xbox Cloud Gaming 让玩家无需主机即可在手机上畅玩高配游戏。国内市场，腾讯 START 云游戏及网易云游戏平台通过边缘节点部署，成功支撑了《黑神话：悟空》等高性能游戏的云端运行，用户无需购买昂贵显卡即可获得 4K、60 帧的极致体验。

此场景对网络时延要求极高（通常需<20ms）。湖北算力互联互通平台（详见附录二）通过优化省内边缘算力布局，显著提升了此类实时交互应用的传输效率，推动了游戏产业向订阅制与流化服务转型。

### 2.云电脑/云手机

云电脑/云手机服务将操作系统封装为云端算力流，实现“终端瘦身，算力上云”。中国电信“天翼云电脑”与中国移动“云手机”已广泛应用于办公与政务领域，企业数据不落地，安全性大幅提升。

在具体实践中，成渝（四川）算力互联互通平台（详见附录五）协同运营模式，让西部地区的教育和企业用户能够以低廉价格通过瘦客户端接入高性能计算环境，实现了算力像水电一样的普惠接入，有效促进了混合办公模式的普及。

### 3.智能体词元服务

智能体词元服务面向模型推理和任务执行等综合能力。服务提供方不生产模型，而是整合并售卖运行这些模型所需的标准化算力（以词元为计量单位），用户通过它购买的是通用了的、可调度的计算效果，而非直接对接某个具体的模型提供商或算力供应商。这是智能算力服务走向互联互通化、互联网化的高级形态。

以 OpenClaw（龙虾）为例，其通过对话入口调用模型和工具能力，支持邮件处理、网页操作等任务。前端体现为对话即服务，后端依赖算力与工具统一编排。随着应用普及，智能算力服务将进一步走向按词元计量、按调用结算的模式。

## （三）垂直行业融合与边缘协同场景

智能算力服务已成为推动行业数字化转型和智能化升级的核心基础。这些场景的典型特征是算力与行业深度融合，且常常需要低时延、本地化处理，形成“云-边-端”一体化的分级协同架构。

### 1.智慧工业

在智慧工业领域，算力服务已从辅助管理走向核心控制。例如，西门子利用边缘智算节点对生产线进行实时优化。在国内，某新能源电池巨头引入“端侧采集+边缘推理”架构进行极片缺陷检测，将缺陷检出率从 90%提升至 99.9%以上，单张图片处理时延降低至 10 毫秒以内。

长三角算力平台（详见附录一）不仅服务于科研，更连接了大量工业互联网节点，支持制造企业利用云端进行模型训练迭代，边缘端

执行毫秒级控制，有效降低了设备意外停机时间，推动制造业从自动化向智能化迈进。

### 2.智慧金融

在智慧金融领域，智能算力服务正加快向业务场景深度渗透，支撑金融机构提升服务效率、风险识别能力和智能化运营水平。例如，在银行网点、智能柜台和移动终端等场景中，智能算力服务可支撑身份核验、票据识别、语音交互、风险提示等业务的实时处理，提升业务办理效率与客户服务体验。

随着普惠金融、智能风控和数字化运营持续推进，智能算力服务将进一步推动金融服务由系统电子化向服务智能化升级，持续增强金融机构的服务能力、风控能力和运营效率。

### 3.智慧零售

在智慧零售领域，智能算力服务通过边缘节点实现“人-货-场”的毫秒级感知与交互。例如，大型商超通过实时分析摄像头采集的客流热力图与货架缺货率，动态调整电子价签与促销策略，压缩营销响应时延。

依托算力互联互通平台，零售企业可将全国门店的销售数据在边缘侧脱敏后汇聚至云端，进行大盘趋势预测与供应链优化，形成“边缘敏捷响应、云端智能决策”的高效协同闭环。

# 六、智能算力服务市场与产业链分析

在前文系统厘清智能算力服务内涵边界、明确三层体系架构与关键支撑技术、梳理多领域典型应用场景的基础上，智能算力服务已从技术探索与场景试点，全面进入规模化发展与全链条竞争的关键阶段。本章将从全球与中国两个维度，深入剖析智能算力服务的市场规模增长态势与竞争格局演变。同时，从产业链全维度出发，解构上游芯片制造、中游服务器与数据中心、下游云计算服务的竞争格局，对比国际博弈焦点与国内发展优势及短板，为后续研判产业发展趋势、提出优化路径提供坚实的数据支撑与现实依据。

## （一）市场规模分析

### 1.全球市场

**从全球市场看，智能算力服务结构日益清晰，进入以“调度效率”为核心的高速增长与生态重构期。**Amazon、Microsoft、Google Cloud、Alibaba、Oracle 占据全球智能算力服务市场份额前列[2]，正在加速将通用云服务向 AI 算力服务转型。由于国外暂无 AI 算力服务的统计口径，因此使用较对应的 AI 算力来表征其市场发展态势。据 Gartner 2025 年发布的《AI 优化服务器发展预测 2023-2029》（《Forecast: AI-Optimized Servers, 2023-2029》）报告数据显示，2025 年全球 AI 优化服务器( AI-optimized Servers )市场支出预计达到 2800 亿美元( 约合人民币 2.03 万亿元），2026 年有望进一步增长至 3530 亿美元，年

[2] 排名不分先后

均增速达到 26.2%。而从更广泛的 AI 基础设施（含服务器、网络、存储等）来看，2026 年全球支出预计将突破 1.36 万亿美元。

市场增长的动力正从基础设施的简单扩容，转向对跨域、异构资源的全局优化配置，此趋势驱动产业生态结构日益清晰，形成三大主导力量内部竞合发展格局：以 NVIDIA、Intel 等芯片厂商为代表的“硬件引领”生态，通过软硬件一体化方案降低调度复杂度，从底层驱动产业创新；以微软、亚马逊、谷歌等云服务商为代表的“纵向整合”生态，通过构建跨域调度体系强化其市场主导地位；以美国电话电报公司（AT&T）、德国电信股份公司（DeutscheTelekom）等电信运营商为代表的“连接赋能”生态，致力于打造低时延、高可靠的全球算力互联网。三方共同推动全球市场在 AI、安全、效率三大核心需求驱动下保持稳定增长。

### 2.中国市场

**从中国市场看，规模扩张与结构优化同步推进，增长势头显著优于全球。**据中国信息通信研究院测算，2025 年我国智能算力服务市场规模预计超 1300 亿元。阿里云、天翼云、移动云、华为云、腾讯云占据我国智能算力服务市场份额前列[3]，市场竞争重心正从基础算力资源服务，向以大模型为核心的模型即服务（Model as a Service，以下简称 MaaS）升级。市场结构呈现“双轮驱动”特征：AI 云企业级服务 800 亿元持续巩固的同时，以云电脑、云电脑等“卡时”消费级市场 500 亿元迅速崛起（其中云电脑突破 1000 万用户，增长率超 100%），

[3] 排名不分先后

成为最活跃的增长极。这一结构性变化背后，是产业各方在政策指引下围绕“算力互联互通”重新定位其市场角色：基础电信运营商强化角色，成为算力“并网”与流通的关键推动者；云服务商积极转向开放平台，接入协同调度体系；设备与解决方案商则致力于解决异构融合挑战。同时，AI 应用向推理侧和终端的扩散，正催生智能体、智能终端服务等新增长点。国家数据局数据显示，2024 年初我国日均词元调用量为 1000 亿，至 2025 年底跃升至 100 万亿，2026 年 3 月已突破 140 万亿，两年增长超千倍。这些数据预示着我国智能算力服务市场正迈向以“网”带“算”、业态创新的高质量发展新阶段。

## （二）产业链分析

智能算力服务竞争正从单一加快演变为覆盖芯片、软硬件、服务器、数据中心、云平台到应用的全链条竞争。美国凭借芯片、云平台和生态协同优势，形成纵向整合能力；欧洲围绕技术主权、绿色算力和可信云加快布局。全球产业由资源规模扩张转向以核心技术、资源组织和生态整合为支撑的体系化竞争。我国在资源供给、设施建设、平台汇聚和融合应用有进展，产业链协同增强，但高端芯片、基础软件、跨域调度、服务生态培育等方面仍需提升。

### 1.上游：芯片、服务器与数据中心

**国际方面，算力芯片领域，呈现“美国主导、欧盟追赶”的二元格局。**美国凭借英伟达、AMD、英特尔等企业在 GPU 和 AI 芯片领域的绝对领先地位，通过《芯片与科学法案》（《CHIPS and Science Act》）强化本土制造，并主导了 NVLink 与 UALink 两大互联技术路线的竞

争。欧盟则通过《芯片法案》（European Chips Act，简称 ECA）和欧洲共同利益重要项目（Important Project of Common European Interest，简称 IPCEI）项目聚焦技术主权，支持英国图芯（Graphcore）、法国思珀尔（SiPearl）等本土设计企业，并积极引进台积电、扩建英飞凌工厂以提升制造产能，力争 2030 年实现全球产能份额 20%的目标。**服务器制造与数据中心建设呈现“中美双核驱动，区域集群化”特征。**美国以戴尔、超微等服务器厂商和谷歌、微软等科技巨头为主导，在弗吉尼亚、俄勒冈、得克萨斯等州形成算力“铁三角”，并通过能源部站点加速 AI 基础设施全国布局。欧盟则以法国源讯（Atos）等本土服务器企业为基础，依托“欧洲高性能计算共同计划”在德、法、西等国建设绿色超算中心，强调可再生能源使用与区域算力枢纽构建。

**国内方面，芯片产业在“依赖与突破”中循序前行。**高端通用 CPU 和 AI 训练芯片仍不同程度依赖英特尔、AMD 和英伟达，但以华为昇腾、海光、飞腾为代表的国产芯片已在政务、特定行业场景取得应用突破。然而，先进制程制造受限、EDA 工具及 IP 核等基础环节稍显薄弱，仍是制约全产业链自主化的核心挑战。**服务器市场成熟且竞争激烈，浪潮、新华三、华为等厂商在全球市场份额居前。**数据中心产业在“东数西算”工程引导下，向西部可再生能源富集区规模化、绿色化转移，并涌现出万国数据、世纪互联等大型服务商。同时，电信运营商和云厂商向下整合，自建或定制数据中心以优化性能成本，形成“软硬件一体化”趋势。

### 2.中游：算力互联互通及调度平台

**国际方面，云计算市场由美业主导但形成生态分化特征。**AWS、Azure、GoogleCloud 三大云巨头合计占据全球超半数份额，其中 AWS 通过自研芯片 Trainium3 和绑定 Anthropic 构建全栈自主生态；Azure 则深化与英伟达的全栈协作强化 AI 超级计算能力；GoogleCloud 则以开放合作策略携手 OpenAI，强化推理服务与区域市场响应。欧盟则通过法国欧维奥云（OVHcloud）的“主权云”和盖亚-X（Gaia-X）联盟，推动数据本地化与可信云基础设施建设，寻求差异化突破。**算力企业积极建设调度平台构建生态壁垒。**2025 年 5 月，英伟达推出的 DGX Cloud Lepton 全球统一英伟达 AI 算力调度与多云管理平台，表明其已不再只是局限于单纯的硬件供给，正转型为全球 AI 算力的“总调度员”。该平台汇聚全球开发者的算力需求，通过统一的算力市场（Marketplace）智能匹配到接入平台的数万颗英伟达 GPU 上（无论这些 GPU 位于 AWS、Azure 还是 CoreWeave 等云厂商机房）。这一模式跨越了底层云平台差异，使开发者无需关心底层基础设施，从而将用户深度锁定在英伟达的 CUDA 生态中。

**国内方面，云计算市场由巨头主导形成“多强争霸”格局。**天翼云、移动云、联通云等国资云，阿里云、华为云、腾讯云等民营资本云等六朵云覆盖主要市场份额。服务模式从 IaaS 红海竞争向提升毛利率的平台即服务（Platform as a Service，以下简称 PaaS）/软件即服务（Software as a Service，SaaS）深化，算力互联互通平台与交易类新业态加速发展，旨在构建全国一体化算力网，降低算力使用门槛、

促进算力资源商品化流通。**市场各方积极推进算力互联互通。**地方政府积极建设算力调度平台，汇聚区域内算力资源形成资源池，支持异构算力的统一接入、统一封装、统一调度。基础电信运营商依托骨干传输网络推进算力并网行动，如移动“算网大脑”、电信“息壤”平台等，纳管社会其他闲散算力或者自身不同类型算力。中国信通院与全球移动通信系统协会（简称“GSMA”）签署开放网关（简称“Open Gateway”）合作备忘录，贡献算力标识网技术，推动全球算力资源标准化感知。基于各方探索基础上，产业界提出全国算力资源“一套标识、一套标准、一套规则”的互联互通需求。

### 3.下游：模型即服务（MaaS）与词元服务

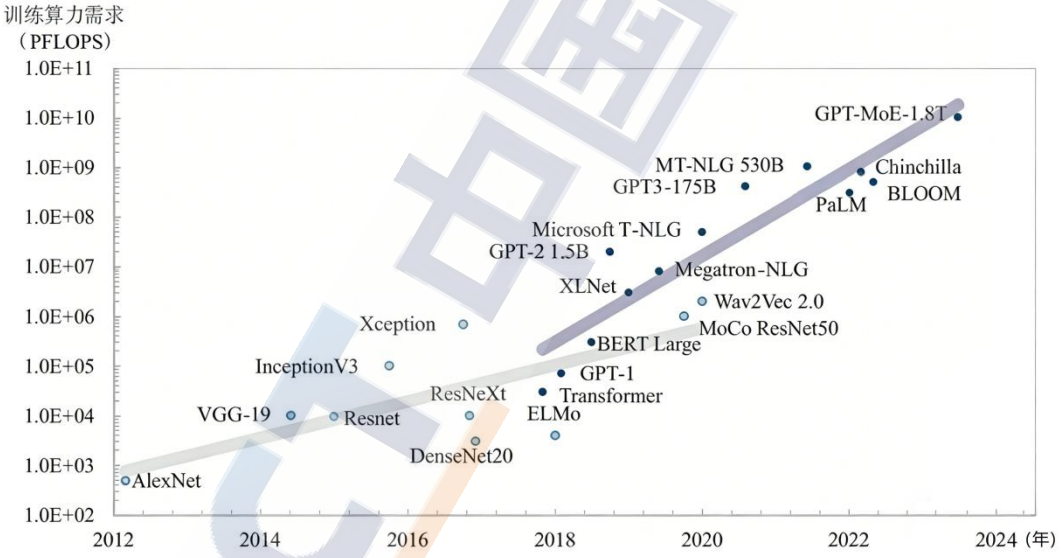
**国际方面，美国头部大模型企业占据市场绝对主导地位，欧盟则聚焦本土合规与技术主权加速追赶。**美国 MaaS 市场呈现头部高度集中、生态深度绑定的特征，主要由 OpenAI（ChatGPT）、Google（Gemini）、Anthropic（Claude）三大核心供给主体掌控。三巨头依托通用大模型与行业大模型双轮驱动，覆盖企业级定制、个人应用、垂直行业解决方案等多元场景，并以词元作为大模型服务计费、算力消耗计量的核心单元，实现全场景标准化计价。从公开数据可见，行业已进入规模化商用爆发期，且头部效应极其显著。根据各家 2026 年初官方最新披露显示，OpenAI 每日词元调用量（API 形式）超 21.6 万亿，Google 则超 14.4 万亿；Anthropic（Claude）具体数据虽未公开，但其与前者共同构成了稳固的美国市场第一梯队。与此同时，欧盟聚焦数据安全与技术主权，积极推动《人工智能法》，筹建欧盟机构级大模型（EU

institutional LLM），明确选用了本土企业法国米斯特拉尔人工智能公司（Mistral AI）的 Mixtral 系列模型作为基础模型进行增强训练。

**国内方面，MaaS 市场呈现“互联网云厂商领跑调用量、运营商云主导政企渗透”的双轨格局。**火山引擎、阿里云、腾讯云、智谱 AI、百度智能云为核心服务商，业务聚焦模型精调、行业专属模型部署、智能体（Agent）服务三大方向。词元作为大模型算力消耗与计费的核心单元，已实现从通用 API 到企业私有化场景的标准化计价。人民日报数据显示，豆包大模型 2026 年 3 月的日均词元调用量已突破 120 万亿，较 2024 年 5 月首次对外发布时增长 1000 倍；智谱 AI（GLM）词元 API 调用量为 4.2 万亿个/日，阿里云（通义千问）、腾讯云（元宝）、百度智能云（文心一言）数据未公开披露。运营商侧，中国电信“星辰”、中国移动“九天”、中国联通“元景”继续作为政务、工业、教育、医疗等行业 MaaS 私有化部署的重要载体。截至 2026 年 3 月，三家运营商均未公开披露统一口径的日均词元调用量数据。

## 七、智能算力服务发展趋势

在产业规模持续扩大、全链竞争加快演进的同时，智能算力服务的需求侧也在发生深刻变化。一方面，人工智能模型算力需求持续攀升，且呈现出由训练主导向训练与推理并重、由中心集聚向分布协同演进的结构性特征；另一方面，用户对智能算力服务的诉求也由获取底层资源，转向获取任务能力、结果交付与普惠化服务。在此基础上，智能算力服务正沿着资源部署、服务交付，产业组织和价值创造四个维度加速演进，呈现出由“资源供给”走向“能力服务”、由“局部建设”走向“全域协同”、由“单点支撑”走向“生态赋能”的发展趋势。




来源：《信息通信技术与政策》2024 年第 6 期

图 8 人工智能模型算力需求

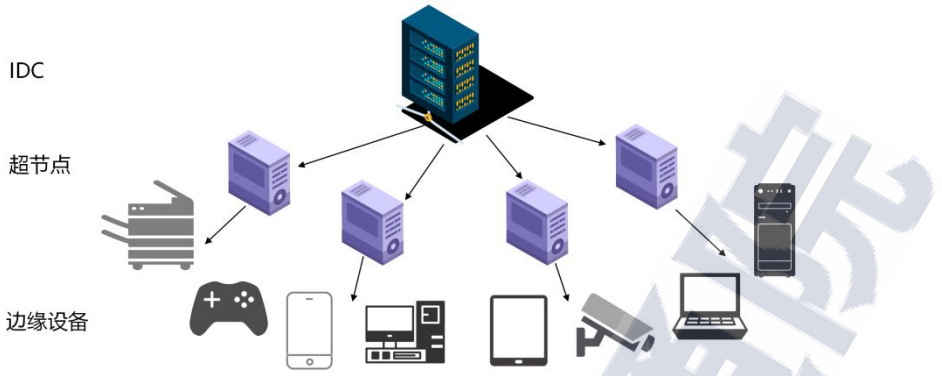
### （一）架构部署：从“中心化”向“云边端高频协同”演变

从需求侧看，人工智能模型算力需求在总量快速增长的同时，正加快由单一扩张向结构分化演进。面向基础大模型训练的集中式高密度算力需求仍在持续增长，要求智能算力服务具备跨数据中心的万卡级调度、高可靠运行和长周期稳定供给能力；与此同时，随着模型应用加快落地，推理侧需求快速释放，呈现出高并发、碎片化、低时延、就近接入等特征，推动算力资源从中心节点向边缘侧、终端侧进一步延伸。训练与推理并行发展的需求结构，决定了智能算力服务不再适合单一中心化部署模式，而是加速走向云、边、端协同的复合型架构。

### 1.中心化向分布式架构的跃迁

从技术架构演进视角看，智能算力服务的物理布局正经历重构，正由 IDC 集中式部署加速向“中心+边缘超节点”的协同分布式架构跃迁，标志着算力基础设施正式进入协同化、场景化发展的新阶段。这一转变的核心是实现能力互补和满足不同业务对时延、带宽及算力密度的差异化需求。具体的部署模式形成了“大中心+小节点”的深度互补范式。

大中心即 IDC 大中心，继续承担海量数据存储、通用计算与批量处理等存量业务，发挥其规模化和成本优势。小节点即边缘“超节点”，作为部署在用户侧或网络边缘的关键设施，提供高敏捷、低时延、高带宽特性，直接赋能实时性业务。例如，在智能制造中，超节点设备可部署于工厂本地，实现工业视觉质检的毫秒级响应；在自动驾驶领域，路侧超节点与云端协同完成高精地图的实时局部更新与全局融合处理。




来源：中国信息通信研究院

图 9 “IDC+超节点”的协同范式

从产业整体发展来看，这种“大中心兜底、超节点敏捷响应”的协同范式已逐步构建起覆盖云、边、端的立体化算力互联网，支撑智慧金融、云 VR/AR 等产业实现业务敏捷部署与服务质量提升。

### 2.互联互通与云边端深度融合

我国智能算力服务正加速向“全国互联与高效传输”的新范式跃迁，其核心在于构建高性能网络和跨域调度平台，实现算力与数据在云、边、端之间的全域、弹性、按需流动。国家算力互联网平台作为统一的调度中枢，通过算力标识体系汇聚了超百 EFLOPS 资源，能够智能路由算力任务和数据流，确保跨区域资源的高效整合。在网络传输层面，以中国电信和华为联合开展的百公里级 AI 广域网试验为代表，通过突破性应用广域 RDMA 技术，成功将跨域训练任务时延压降至惊人的 1.1 毫秒，极大地提升了大规模计算节点间的数据传输效率，不再依赖 CPU 介入，是支撑云间高效协同的关键。

同时，网络传输和调度也支撑了“云-边-端”的深度融合与分级协同。随着云电脑、云手机等形态的普及，智能算力服务和数据流实现

了按需供给，终端能力由云端弹性提供。在自动驾驶和智慧工业等高频互动场景中，数据流必须在负责模型训练的云端、负责实时推理的边缘，以及负责感知交互的终端之间，进行双向、低时延的闭环传输。这一体系依赖于网络具备分级传输和智能路由的能力，确保关键数据以最高可靠性和最低时延在不同层级间流动，从而驱动产业智能化升级。

## （二）服务模式：从“资源供给”向“任务式交付”升级

随着算力应用门槛持续降低，用户对于智能算力服务的需求正在由购买资源转向调用能力、由租赁硬件转向获取任务结果。与传统算力服务相比，当前市场更加关注服务的标准化、普惠化与精益化：一方面，用户希望像使用通信流量一样便捷获取算力服务；另一方面，云渲染、云游戏、云电脑、智能体调用等场景又要求服务商能够面向任务自动完成资源匹配、跨域调度与统一结算。由此，智能算力服务模式正从“资源供给”加快走向“任务式交付”。

### 1.“任务”为中心的交付模式

智能算力服务模式正在经历从“资源租赁”向以“任务式服务”为核心的“结果交付”的深刻变革。这种转变的核心在于屏蔽底层复杂性和实现交付标准的细化。

技术上看，服务商通过构建智能化的资源调度系统、API 或流化技术，将 CPU、GPU、DPU、FPGA 等底层异构算力进行统一封装。用户无需自行管理庞大的集群或复杂的配置，而是直接面向业务逻辑或最终结果。这种以“任务”为中心的交付模式，通过动态调度最优算

力资源，极大地提升了算力的普惠性和利用效率。

计费与交付标准也从“资源规格+时长”转向“任务复杂度+结果”。例如，在云渲染与 AI 模型服务中，用户直接购买的是“渲染作业的完成”或“AI 能力的调用次数”。系统后台自动进行资源的动态调节与精准匹配，真正实现了按需取用。

云游戏是任务式交付模式的典型代表。腾讯 START 云游戏、网易云游戏等平台通过直接交付游戏的运行结果流，成功支撑了《黑神话：悟空》等高性能游戏的云游戏体验，打破了用户本地硬件瓶颈，充分证明了任务式服务在支撑高性能计算需求方面的产业价值和广泛的市场接受度。

### 2.驱动“云–端–智”高效协作

终端载体变革是当前智能算力服务对用户体验最大的颠覆。云电脑、云手机等新型云终端的普及，标志着智能算力服务正加速实现“前端轻载、后端强算”的高效协作模式。用户手中的设备不再需要内置昂贵的、强大的计算硬件，而是演变为轻薄、灵活的智能服务触点。这种转变极大地优化了用户体验：设备形态更轻便、成本更低，但用户可调用的算力却通过云端实现弹性调度和按需供给，性能远超任何单体设备。例如，工业工程师仅需一个轻量化的云电脑，就能远程调用云端高性能算力，轻松完成复杂的产品仿真与设计，使得专业级任务的普惠性和移动性大幅提升。

这种“云–端–智”的深度融合，正在加速智能算力服务的互联网化和个性化。它支撑着亿级用户规模的移动办公、云游戏等场景，用户

体验不再受限于设备的硬件配置，而是能享受到多端协同带来的无缝流畅。通过将核心算力集中于云端，并以轻量化终端为载体，智能算力服务正实现场景化、个性化与普惠化的纵深发展，让强大的计算能力成为一项随时可取用的公共服务，彻底模糊了物理终端与数字世界的界限。

## （三）产业格局：从“独立发展”向“算力互联网”聚合

训练与推理需求分化、跨域调度需求上升以及服务交付向标准化、任务化演进，共同推动智能算力服务产业从分散建设、独立运营走向互联互通、平台聚合和规则统一，算力互联网正成为资源高效流通和价值再组织的重要载体。

### 1.产业从孤立走向协同，以平台化架构支撑算力高效流通

智能算力服务已超越基础供给，成为驱动人工智能规模化应用、赋能千行百业的核心载体。大模型训练等前沿 AI 任务对万卡规模集群和跨数据中心协同调度提出了刚性需求，这推动了算力互联互通平台的技术架构发展，旨在实现“跨主体、跨地域、跨架构”算力高效协同。与此同时，智能算力服务产业正走出各自为战的孤岛，通过标准化与互联互通手段，加速形成开放流通的交易生态。

### 2.交易从无序走向标准化，以普惠服务降低 AI 应用门槛

智能算力服务核心之一在于算力交易与调度标准化：依托“算力互联互通平台”等中枢设施，产业界正致力于构建统一的算力度量、算力标识与交易结算体系。这使得不同归属方的算力资源能够像商品

一样在平台体系上进行查询、竞价与流转，解决了跨地域、跨主体算力利用难的问题。这种生态模式的互联网化使智能算力服务演变成一个包含算力资源服务方、算力互联互通服务方、算力应用服务方等多元主体的复杂生态系统，各方通过标准接口实现能力的互通与能力的原子化封装，共同推动智能算力服务向更高阶的互联网服务模式升级。面向更广泛的市场，智能算力服务正通过智能化赋能实现普惠化，例如运营商推出的“算力卡”、“云电脑”等服务，本质是将强大的 AI 算力以任务化、普惠化的方式提供给个体用户或中小企业，显著降低了其拥抱 AI 的技术与成本门槛，从而加速整个社会的智能化转型进程。

## （四）赋能路径：从“计算能力”向“生态价值”延展

随着智能算力服务由资源供给、任务交付进一步走向生态协同，其价值释放方式也在发生深刻变化。智能算力服务不再局限于提供可调用的计算能力，而是通过与模型、智能体、终端应用及行业场景深度融合，逐步演化为支撑多主体协同创新、促进普惠化用算、赋能产业智能升级的重要基础性能力。

### 1.支撑多智能体协作，促进智能体互联互通

算力互联网作为智能体协同与进化的“能力基座”，通过“智能感知，实时发现，随需获取”三大机制，为多元智能体提供标准化、服务化的算力供给。这打破了智能体之间的“孤岛”状态，促进智能体互联互通，最终形成具有更高复杂性和适应性的智能体生态系统。作为底层支撑，算力互联网整合和灵活调度算力资源，满足多智能体系统在复杂任务中的高并发需求和算力激增的挑战，从而支撑多智能体在

复杂场景中规模化落地。

智能体互联互通代表了从单一智能体到群体智能生态的跨越式发展。这包含互联和互通两个层面的要求。互联属于物理/逻辑层面的“连接”，涉及通信协议、接口规范等，旨在使不同智能体之间能够传输数据和信号。而互通则属于语义/功能层面的“理解与协作”，涉及语义互操作性、行为协同性、以及安全与信任机制等，目的是使智能体之间能够理解彼此的意图并有效协作。算力互联网的支撑层级从底部的云、IDC、通算、智算、超算等资源，向上通过编排调度，提供 MaaS 和智能体即服务（Agent as a Service，简称 AaaS），最终实现智能体互联互通。

### 2.普惠化用算，推动人人皆可用的数字化能力

智能算力服务正成为社会基础设施的一部分，像水、电、网络一样。它以低成本、高效率的方式向每个人、每个组织提供算力资源，推动教育、医疗、交通等各领域的数字化升级。实现普惠化用算具备三大核心特征：全民触达，即打破技术壁垒，让不同人群都能便捷使用算力；低成本获取，通过租赁、共享等模式降低使用门槛，使用户无需进行重资产投入；以及全场景覆盖，将智能算力服务从个人生活、企业生产延伸至城市和乡村，渗透至千行百业。

普惠算力的应用体现在三大场景：个人数字生活助手，如支持 AI 个人助手、元宇宙社交、智能创作等，用户可通过手机、家庭终端一键调用弹性算力；小微企业“轻量算力包”，企业可通过算力超市、共享平台按需租赁 AI 客服、智能财税等服务；以及社区智慧服务中

枢，利用边缘算力节点支撑社区安防、养老看护、便民服务等实时响应的智慧化服务。

### 3.赋能新型工业化，智能制造新引擎

面向新型工业化，算力互联网将全面赋能制造业、能源业、交通业等传统产业，推动产业智能化转型，助力实现绿色、低碳、高效的工业体系。算力赋能新型工业化主要体现在三大典型应用场景：在智能工厂中，企业无需大规模自建基础设施，可按需调用算力互联网的算力；在数字孪生中，依托算力互联网的弹性供给，企业无需长期投入海量资源，即可低成本、灵活地构建虚拟工厂；在仿真计算中，企业无需常备大规模高性能算力设施，在开展产品设计、工艺仿真等复杂任务时可随时便捷调用所需算力。

算力赋能新型工业化将带来核心效益：高效生产，通过云计算和边缘计算协同提升生产效率，实现柔性制造和精准控制；绿色低碳，基于智能算力调度和优化算法，有效降低工业能耗，推动碳中和；智能决策，依托强大的算力支持，实时分析工业数据并进行仿真计算，从而提升全链条的生产效能。

# 附件：智能算力服务的典型场景

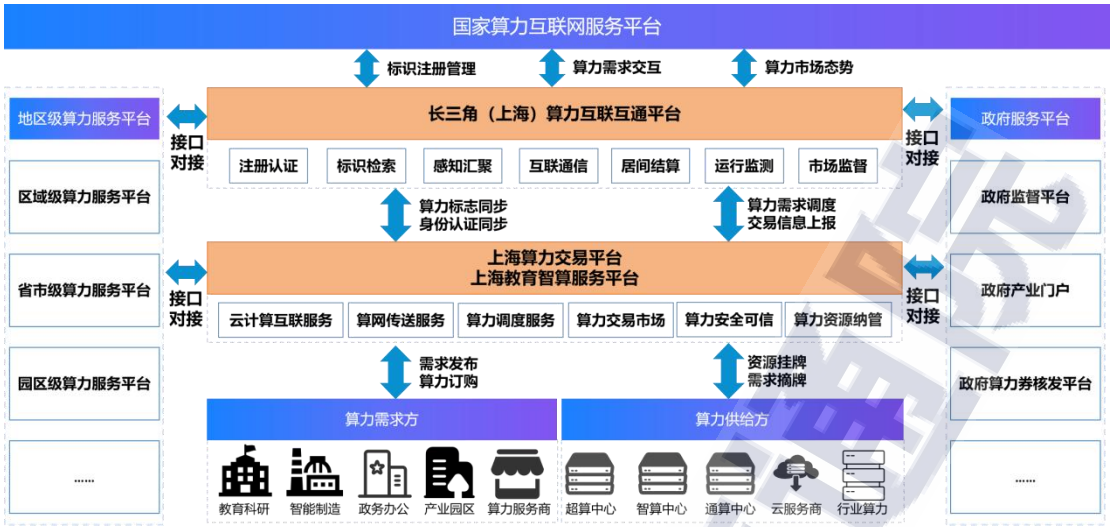
## （一）长三角（上海）算力互联互通平台实践

### 1.案例背景

长三角区域教育与科研领域的算力需求旺盛，但现有资源分布零散，缺乏统一的调度标准，导致跨区域协作难、资源复用率低。为打破孤岛效应，亟需建设区域级平台，通过统一标准实现算力资源的集约化共享。

### 2.解决方案

一是专项行动计划，发布“算力沪联”专项行动，明确未来 3 年的推进任务；二是算力应用创新试点，率先在智慧教育、智慧医疗等民生领域探索算力互联互通体系在应用上的赋能创新；三是算力互联互通平台，建设长三角（上海）算力互联互通平台，实现与国家平台和各个算力调度平台的互联；四是算力运行安全监测，以算力运行安全的强监管为依托，推动各个算力中心纳入算力互联互通体系；五是算力调度合规体系，以新业务备案的方式，推动算力调度业务平台的合规化运营，强化算力接入纳入算力互联互通体系。




来源：中国信息通信研究院

图 10 国家算力互联网服务平台架构图

### 3.服务模式

依托国家（上海）新型互联网交换中心构建的算力专网、上海算力交易平台作为算力算网供给底座，通过打造教育教学应用平台、高校科研实训平台，将教育行业模型应用生产企业、基础教育学校、高等院校、科研院所高效衔接起来，为上海、长三角区域乃至全国的各类院校在开展教育实践重塑、人工智能教育课程建设、科技研发和学术研究等方面提供完整的服务和赋能方案。

### 4.赋能成效

成本效能方面，项目通过纳入算力互联互通体系，向下应用了算力标识体系对多元算力实现了归一化管理，向上为教育行业实现了算力统一化供给服务，通过集约化的资源共享模式减少重复投资和资源浪费，降低设备和服务的采购成本，提高资源的利用效率。

业务模式方面，通过本次创建和推广示范案例和成功模式，推动

人工智能教育教学模式的创新。新型数智化技术赋能的教育应用场景将为教育机构提供新的发展模式，并引导教育行业向数字化和智能化方向发展。

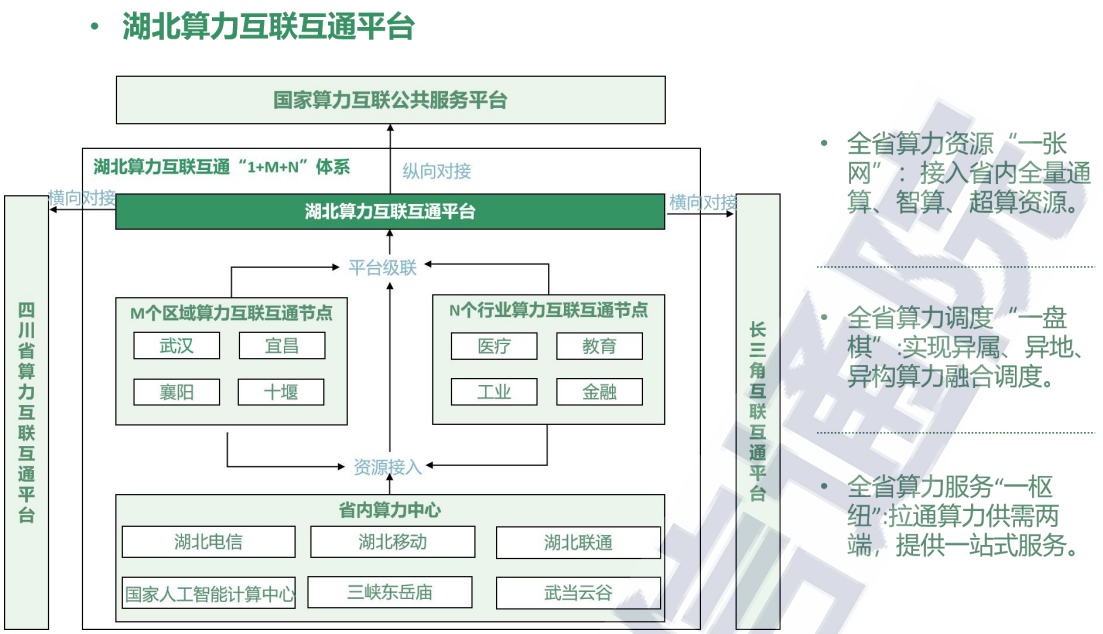
## （二）湖北算力互联互通平台运营实践

### 1.案例背景

湖北正致力于打造全国算力中部枢纽，但当前面临着算力基础设施布局不均、“云-边-端”协同不足的问题，由于缺乏统一的市场化调度机制，大量优质算力未能转化为服务效能。为此，构建“1+M+N”的省域互联互通体系，旨在物理上连接孤立的资源池，逻辑上形成全省统一算力市场，从而优化资源配置，提升产业赋能水平。

### 2.解决方案

湖北算力互联互通平台是遵循国家相关政策标准，以全省算力统一纳管、融合调度、便捷服务、产业赋能为目标，打造的重点数字基础设施，是全省算力资源调度与服务中心。平台定位算力互联互通中部区域节点，是湖北打造全国算力互联网中部枢纽的重要抓手。平台通过对接全省算力中心，高质量整合全省算力资源，具有中立开放、智能高效的特点，旨在优化区域算力供需配置，提升省内算力应用水平。




来源：中国信息通信研究院

图 11 湖北算力互联互通体系总体架构

## 3.服务模式

湖北算力互联互通平台围绕多元化的算力服务需求，构建了覆盖“算力—增值—生态”三位一体的综合服务体系。在算力服务方面，平台提供多种形式的算力供给模式：包括资源型服务，支持线上云资源的开通与全生命周期管理，以及线下智算、超算服务器或芯片的租赁与销售；任务型服务，允许用户直接提交大模型训推、科学计算、模拟仿真等专业化计算任务，由智能调度引擎自动匹配全域资源，并提供成本、时长、安全、网络时延、绿电等多种优化策略；平台型服务，则通过算力互联互通与调度能力，整合省内外资源形成统一虚拟资源池，支持中训边推、异构跨域混训混推等协同模式，并提供包括大模型训推平台、科学计算平台等 PaaS 服务，以及按词元计费的大模型推理 API；此外，平台还提供算力调度服务，开放北向接口供第三方

平台接入，共享全域算力资源与调度能力。

在算力服务基础上，平台进一步提供增值服务，围绕大模型研发应用、科学研究、视频渲染等具体场景，整合数据、算法、模型、软件、网络、存储、安全等要素，为客户提供一站式、全链路的解决方案。

同时，平台着力构建生态服务，在数据、算法、模型、应用软件、网络、存储、安全等领域建立服务商体系，并为生态伙伴提供供需对接、市场情报、政策解读、投融资、人才培育、成果转化等专业化支持，从而形成从算力供给到场景支撑、从资源整合到生态协同的全方位服务闭环。

#### 4.赋能成效

通过高水平建设全省算力资源“一张网”、算力调度“一盘棋”和算力服务“一枢纽”，着力打造一站式算力服务、加速培育算力产业链生态、全面融入全国算力统一大市场，盘活省内千亿级算力市场。平台建成后，以算力为抓手，能够充分发挥我省产业、人才优势，实现数据、算法、应用场景等要素的融通利用，通过构建算、数、产深度融合的服务体系，形成“以算育数、以数育产”的发展新格局，以“AI+”提升湖北产业发展能级，为湖北实现“51020”战略提供数智动能。

### （三）深圳市算力互联网建设运营及赋能实践

#### 1.案例背景

深圳市人工智能产业的算力需求持续增长。然而，当前其智能算力资源稍显匮乏且分布零散，未能形成有效的整体效能。这种碎片化

现状难以支撑大型模型训练与推理所需的算力。因此，构建城市级的算力互联网，实现算力资源的全面整合与优化，加速 AI 创新应用的落地与发展，已成为深圳市亟待解决的关键问题。

### 2.解决方案

深圳市建设了“深圳开放智算中心”与“深圳市智慧城市算力统筹调度平台”，旨在打造城市算力互联网基础设施平台。中心位于深圳河套深港科技合作区的 5A 级机房内，依托弹性伸缩、分布式的云底座，实现了底层硬件和管理网络的统筹管理。同时，平台基于深圳 SDN 新型城域网和广东云网汇接中心网络能力，集算力建设、算网融合、算力调度于一体，不仅服务于深圳本地，更计划拓展至粤港澳大湾区乃至全国和全球。

中心搭载了业内主流的 AI 开放框架，并采用了 Mediated Passthrough (vGPU)和直通透传虚拟化技术、异构算力资源混合调度、以及训推一体任务通义调度等领先技术。这些技术使得算力资源具备高弹性、高性能、高速互联和高性价比等优势，适用于 AIGC 大模型训练、智能推荐、无人驾驶、生命科学等多种场景，为党政军、科研院校、企业等用户提供高性能、灵活化、差异化的智能算力服务。

### 3.服务模式

深圳市智慧城市算力统筹调度平台汇聚了全国 130 多个资源池，集结了超过 17000P 的智算算力、500P 的通算算力和 2P 的超算算力。平台立足深圳，致力于为深圳及全国企业提供多元化的智能算力服务选择。

在用户调用云主机、容器、超算算力、智算任务等资源时，算力调度平台通过智能化、动态化、自动化的推荐策略和任务规划，综合考虑算力需求、网络需求、地理位置、算力规格、算力性能和算力成本等多种维度，结合算网感知和意图感知等能力。通过多维算网资源的联合优化下发，实现了全域资源的创建和最佳分配，有效解决了算力在地区分布不均衡和资源高效使用的问题。

来源：天翼云科技有限公司

图 12 深圳市智慧城市算力统筹调度平台架构图

## 4.赋能成效

“一中心+一平台”的组合模式已彰显出显著的社会效益和行业效益。从社会效益看，平台已成功整合并管理超过 17000P 的算力资源，并预计到 2025 年底可调度算力规模将增长至超过 40000P。平台以深圳开放智算中心构建的算力走廊为基础，已实现向北、西、南（经香港）的跨区域拓展，进一步接轨国际网络。

从行业效益来看，平台为多个关键领域注入了发展动力。在司法审判领域，平台成功赋能全国首个针对司法审判垂直领域的大模型，为深圳市中级人民法院的人工智能辅助审判系统提供了较为坚实的

算力支持。平台系统在 48 类常见诉讼材料的识别提取上准确率超过 95%，极大地提升了司法审判的效率与准确性。此外，平台通过普惠智能算力服务，助力港科大孵化的奇点实验室等初创企业，通过调度贵州节点的算力资源，成功实现了大模型训练目标，推动了区域经济的创新与繁荣。

### （四）中国联通“CubeAI”算网模一体化服务平台实践

#### 1.案例背景

随着 DeepSeek 等国产大模型的爆发，智算需求急剧增长。然而，底层硬件（芯片）的异构差异与上层复杂的业务场景之间存在巨大鸿沟，导致算力服务供需链割裂，调度效率低下。如何屏蔽异构硬件差异、实现跨层跨域的协同调度，成为当前亟需解决的技术瓶颈。本案例旨在构建一体化平台，解决算力割裂痛点，实现 AI 模型的高效服务化落地。

#### 2.解决方案

本案例以“算-网-模一体化协同”为目标，针对从算网资源到行业应用的供需链割裂问题，算网模一体化构建了包括算网资源、算网服务、原子能力、业务场景的四层体系架构，为用户提供跨域动态时变环境下的算-网-模一体协同调度能力，实现了从算力资源互联到算力服务供给的演进。

方案构建了覆盖全层级的全流程智算服务体系。一是知识驱动的闭环模型协同技术与平台构建，适配国产智算，提升模型迁移效率，融合大小模型能力形成闭环；二是算力标识与任播地址结合的算网调

度机制，实现端到端 AI 服务调度；三是构建多领域原子能力模型，并在车联网、智能运维等场景落地应用，推动算网与 AI 融合技术落地。

来源：联通云数据有限公司

图 13 全流程智算服务体系架构图

### 3.服务模式

在资源整合方面，屏蔽异构算力差异，将 x86、Arm、AI 芯片等各类算力统一纳入资源池，实现跨平台协同调度，提升利用率；在应用开发与部署方面，提供一键式服务化封装，支持“代码即服务”，简化开发至部署全流程，提升效率；算力标识与调度方面，采用任播地址+SRv6 技术，具备实时服务发现、动态路径编排及 QoS 保障能力，响应更快、可靠性更高。AI 赋能方面，创新采用大小模型协同机制，在故障诊断、资源调度等场景中大幅提升效率。

### 4.赋能成效

CubeAI 是驱动中国联通向 AI 原生服务商转型的核心引擎，已成

为下一代互联网创新体系的关键组成部分。通过对云、边、端异构算力资源的统一纳管与智能调度，显著提升了算力基础设施的整体利用效率，平台支持主流国产算力芯片，这些实践为构建自主可控的高效算力网络提供了关键技术路径。本案例将技术能力转化为直接的经济效益与市场竞争力，已产生并带动经济效益超过 3000 万元，通过提供模型到智能体的全生命周期管理与服务化封装，CubeAI 降低了 AI 应用门槛，赋能内部业务创新与外部产业合作，支撑“科创中国算力网络创新基地”建设，在多个单位示范应用，助力行业在算力服务、车联网等战略领域开拓新市场、打造新生态。

### （五）成渝（四川）算力互联互通平台协同运营实践

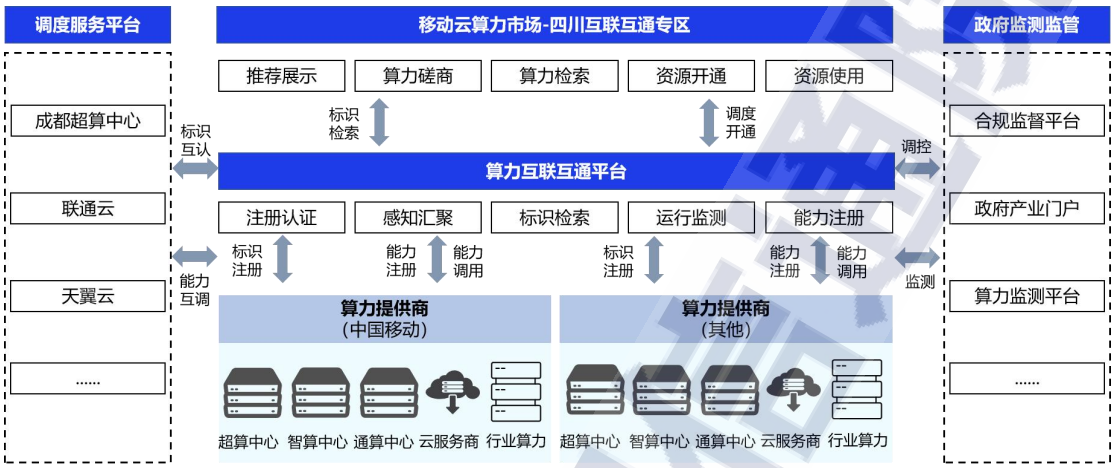
#### 1.案例背景

作为“东数西算”的关键枢纽，四川省内虽拥有丰富的算力资源，但长期分属于基础电信运营商、公有云厂商等不同主体，资源感知与调度能力互不相通。为打破主体壁垒，盘活全省存量算力，亟需建立统一的标识与监测体系。通过算力互联互通平台与运营商“算网大脑”的对接，实现全省算力资源的标准化互认与跨主体协同调度。

#### 2.解决方案

按照“分步实施、迭代演进”的建设思路，开展浅接入、深接入、服务调度三方面的工作。首先，通过全盘梳理中国移动四川省内算力，建立统一算力标识体系，并通过部署算力互联互通网关实现省内算力全面感知、统一监测；其次，进一步标识可调度算力范围，并依托算力互联互通平台注册算力管理能力，实现算力标识跨厂商共享、互认；

最后基于移动云运营门户打造算力市场四川互联互通专区与成渝（四川）互联互通平台开展协同运营试点，通过获取算力标识上架算力产品、调用算力管理能力完成资源开通。




来源：中国移动通信集团公司

图 14 成渝（四川）算力互联互通体系架构图

### 3.服务模式

依托成渝（四川）互联互通平台实现跨算力服务商算力感知和算力互认，实现区域算力整合，打造四川算力底座。并通过算力互联互通平台实现异构算力管理能力统一适配，支撑上层算力调度服务平台统一调度开通。基于移动云运营门户创新打造算力市场四川互联互通专区，开展与成渝（四川）互联互通平台协同运营，支撑跨服务商算力协同运营和一体化服务。

### 4.赋能成效

在算力接入方面，基于统一标识体系实现中国移动四川区域超 540PFLOPS 算力统一接入，支持算力规模、算力使用情况动态监测；同时实现公有云算力的深度接入，支持算力资源开通、退订等深度管

理，支持跨服务商算力互查和互操作。在算力协同运营方面，全面上线算力市场四川互联互通专区，上架超 20 种算力规格，支撑四川安吉物流集团等 10 余家客户在线订购。

CAICT 中国信通院

中国信息通信研究院 云计算与数字化研究所

地址：北京市海淀区花园北路 52 号

邮编：100191

电话：010-62302074

传真：010-62302074

网址：www.caict.ac.cn