

面向词元（Token）服务的智 算基础设施发展研究报告

（2026 年）

中国信息通信研究院产业与规划研究所

2026年9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、摘编或利用其他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，编者
将追究其相关法律责任。

前 言

当前，人工智能正加速从技术突破迈向规模化应用的新阶段，智算基础设施的需求结构随之发生转变，推理算力首次超越训练算力成为需求主体，Token（词元）取代 FLOPS、GPU 卡数成为衡量 AI 产业价值的新标尺。然而，智算基础设施从“算力工厂”到“Token 工厂”并非概念的简单更迭，而是计算范式、技术栈、商业模式的系统性重构。实践中，各方对这一重构的内涵边界、演进规律、推进路径的认识仍不充分，亟须构建系统化的分析框架，以厘清面向 Token 服务的智算基础设施内涵特征，总结差异化的区域布局与产业模式，助力智算基础设施真正实现从“提供裸算力”向“交付 Token 服务”的价值跃迁。

作为智算基础设施系列研究的延续与深化，本报告在《智算基础设施发展研究报告（2024 年）》《人工智能算力基础设施赋能研究报告（2025 年）》基础上，持续跟踪产业演进动态，聚焦面向 Token 服务的智算基础设施的态势特征、政策导向、区域布局、技术突破及产业生态，阐述最新发展趋势与典型实践，致力于为行业企业及相关机构提供决策参考。

态势研判方面，报告系统梳理智算基础设施从“算力工厂”向“Token 工厂”的范式重构进程。人工智能产业的核心驱动力正从模型训练转向推理服务，传统以峰值算力（FLOPS）为价值标尺的“算力工厂”模式，难以满足推理时代对低延迟、高吞吐、弹性调度的要求。新的范式以 Token 为计量单元、以每瓦 Token 产出为核心效率指标，正在重塑硬件选型、集群架构、调度系统与商业模式。

政策布局方面，报告深入分析国家推动构建 Token 友好智算基建体系的政策导向。“人工智能+”行动聚焦智能算力统筹、供需匹配与运营模式创新，全国一体化算力网需适配 Token 级调度与跨域协同，地方政策试点正从“建设补贴”转向“Token 服务采购”，政策侧重点从建设规模转向服务效能。

区域布局方面，报告系统研究智算资源空间配置逻辑的深刻转变。训练任务高度集中而推理任务广泛分布，推动算力布局从单纯追随能源禀赋转向贴合 Token 需求分布，训练算力向枢纽集聚、推理算力靠近需求的分层格局加快形成，枢纽节点与非枢纽区域差异化定位、梯度协同，跨域调度为全国 Token 服务资源优化配置提供支撑。

技术突破方面，报告聚焦提升每瓦 Token 产出的核心能力。面向 Token 服务，智算基础设施技术体系的重心正从“峰值算力最大化”全面转向“每瓦 Token 产出最大化”，核心技术覆盖推理芯片、推理引擎与模型压缩、服务器与集群架构、网络互联与调度系统全栈维度，硬件升级与软件优化协同、架构重构与调度提升并举，推动智算基础设施从“够用”走向“好用”，实现产出效率的持续跃升。

产业生态方面，报告系统梳理 Token 经济驱动下智算产业生态的迭代升级路径。基础设施建设运营从工地建造走向工厂化生产，产业主体格局从资源租赁向 Token 服务运营深度转型，算电协同、分布式协同、算模协同、政产学研用协同等模式共同推动智算基础设施从资源供给走向价值交付，为智算基础设施落地赋能筑牢产业支撑。

当然，报告仍有诸多不足，恳请各界批评指正。

目 录

一、智算基础设施发展态势.....	1
（一）智算基础设施内涵持续演进.....	1
（二）“词元工厂”成为行业关注热点.....	2
（三）Token 服务重塑技术栈与商业逻辑.....	4
二、智算基础设施政策导向.....	5
（一）“人工智能+”行动：强化算力统筹与创新运营服务.....	6
（二）全国一体化算力网：面向 Token 服务的算力跨域协同持续推进.....	8
（三）地方政策试点：突出智算基础设施运营服务效能.....	9
三、智算基础设施区域布局优化.....	11
（一）布局逻辑重构：训练与推理的空间分化倒逼资源配置范式转变.....	11
（二）算力枢纽节点：从“东数西算”到“训推分层协同”的分工深化.....	13
（三）非枢纽节点区域：三级梯度布局支撑 Token 服务广泛落地.....	14
（四）跨区域协同：强化 Token 供需资源优化配置.....	16
四、智算基础设施技术创新突破.....	17
（一）芯片：异构协同降低 Token 生成成本.....	18
（二）软件栈：智算软件体系加速重构升级.....	19
（三）集群：超节点与训推一体重构部署形态.....	22
（四）互联：跨集群协同推进 Token 级调度.....	23
五、智算基础设施产业生态迭代提速.....	25
（一）基建范式：智算基础设施建设运营的全链条升级.....	26
（二）产业主体：三类力量引领智算基础设施角色转型.....	28
（三）生态协作：三类模式引领智算基础设施协同创新.....	31
六、展望与发展建议.....	34
（一）完善政策引导：强化规划协同与服务效能导向.....	34
（二）优化区域布局：推动算力资源分级配置与跨域协同.....	35
（三）夯实技术创新：强化面向 Token 服务的软硬件基础.....	36
（四）培育产业生态：构建多元主体协同的 Token 服务体系.....	38

一、智算基础设施发展态势

人工智能产业的核心驱动力正从模型训练转向推理服务，Token（词元）作为 AI 交互的最小语义单元，正成为算力服务统一的计量与计费单位。Token 服务的兴起，标志着算力服务从非标资源租赁迈向标准化、规模化供给的新阶段，正在重塑智算基础设施的技术栈与商业逻辑。随着人工智能产业发展重心的迁移，智算基础设施的内涵持续演进迭代，从早期以满足通用计算需求为主，到面向大模型训练构建大规模算力集群，再到当前以 Token 为价值尺度、面向推理服务重构技术体系与运营模式，其功能定位、核心能力与价值交付方式均在发生系统性变革。

（一）智算基础设施内涵持续演进

智算基础设施的概念在产业实践中持续演进。2024 年，中国信息通信研究院将其定义为“以智能计算设施为核心，以智能算力集群为载体，面向人工智能应用场景提供算力、数据和算法服务的公共算力基础设施”。该定义确立了计算、网络、存储三维协同的基本框架。2025 年，中国信息通信研究院进一步强调智算基础设施正从基础算力供给向全面支持异构资源智能管理、大规模集群高效调度演进。工信部等十一部门将智算中心界定为“采用人工智能计算架构，提供人工智能应用所需算力服务、数据服务和算法服务的一类算力基础设施”。

综合产业实践演进，本报告将智算基础设施界定为：以人工智能专用算力芯片及异构加速芯片为算力底座，以智能算力集群为核心载

体，以 Token 为统一计量与交付单元，面向 AI 推理与训练场景，提供可度量、可计费、可保障的 Token 生成、调度与交付服务的公共算力基础设施。与传统智算基础设施定义相比，本定义的突破在于：（1）将 Token 确立为核心计量单元，取代传统的 FLOPS 或卡时；（2）将“Token 生成与交付”作为核心服务内涵，而非泛化的“算力服务”；（3）强调服务的可度量性与可保障性（SLA），体现从资源交付向价值交付的转变。这一新内涵要求智算基础设施在硬件选型、集群架构、调度系统和商业模式等全产业链环节进行系统性重构。

（二）“词元工厂”成为行业关注热点

智算基础设施的价值标尺正从峰值算力切换为 Token 产出能力。据 TrendForce 数据显示，2026 年北美五大云端服务供应商的总 AI 训练算力将年增 56% 以上，而 AI 推理算力将年增约 122%，推理算力增速是训练的两倍以上。智算基础设施产业竞争的焦点正从“拥有多少算力”转向“产生多少 Token”。传统以 FLOPS 衡量算力规模、按卡时出租 GPU 的模式，已难以满足推理时代对毫秒级响应、高并发 Token 生成与弹性调度的要求，Token 工厂也顺势成为业界布局重点。

Token 工厂作为智算基础设施的新形态，正在全国加速落地。Token 工厂，是指以 Token 为统一计量单元，面向 AI 推理服务需求，将电力与算力高效转化为标准化 Token 输出，并按 Token 消耗量进行计费与交付的智算基础设施新形态。**算力枢纽区域：**内蒙古和林格尔新区依托中国移动智算中心与火山引擎，联合建成全国最大规模的机密 Token 工厂；甘肃庆阳正全力打造面向全球的“词元工厂”，截至

2025 年底智算规模达 11.4 万 P；广东韶关发布粤港澳大湾区公共算力服务平台及 Token 工厂，已形成 5 个万卡集群。**非算力枢纽区域：**江苏常州签约首个城市级绿色词元工厂，由七方共建，建成后预计年产 Token 60 万亿个；福建厦门启动全省首个城市级全栈 Token 工厂，采用“前店后厂”模式，依托“九州”算力光网实现全国 20 毫秒超低时延；四川雅安上线全省首个 Token 工厂“天绛·白泽 Token 工厂”，年内目标月产能 10 万亿词元；河南郑州启动“杰创郑州词元工厂”项目，面向中部地区输出标准化 Token 服务。

与传统“算力工厂”相比，“Token 工厂”的投入产出逻辑发生了根本变化。输入电力与算力，输出标准化 Token，核心效益指标从总算力规模、上架率转向每瓦 Token 产出效率、单位 Token 成本与 Token 服务质量。英伟达在 2026 年 GTC 大会上提出数据中心的度量标准正在发生变化，“每瓦 Token 产出效率（Token per Watt）”正取代传统的峰值算力与单卡性能，成为新的核心 KPI，另外英伟达 Vera Rubin 平台在每瓦特 Token 生成能力上较上一代提升了 10 倍。商汤大装置率先提出 TPW（Tokens Per Watt）指标，以单位电力成本创造 Token 产出衡量 AI 生产效率，其算电协同 Agent 已在商汤临港 AIDC 落地应用，致力于提升单位电力成本的 Token 产出。百度智能云将 MaaS 模型服务全面升级为“Token Factory 词元工厂”，聚焦于建设每瓦性能更强、性价比更高的 AI 基础设施。

表 1 算力工厂与 Token 工厂范式对比

对比维度	传统算力工厂	Token 工厂
核心价值标尺	峰值算力（FLOPS）、GPU 卡数	Token 产出量、每瓦 Token 效率
投入产出逻辑	输入电力与硬件，输出算力资源	输入电力与算力，输出标准化 Token
核心效益指标	总算力规模、机柜上架率	每瓦 Token 产出、单位 Token 成本、服务 SLA
计费交付模式	按卡时 / 机柜 / 算力规模租赁	按 Token 消耗量计费、结果化交付
调度粒度	任务级、集群级	Token 级、请求级
核心适配场景	大模型训练、离线批量计算	实时推理、高并发交互场景

（三）Token 服务重塑技术栈与商业逻辑

Token 服务需支持毫秒级响应、高并发 Token 生成、动态负载均衡及训推一体化的灵活调度，正深刻改变着智算基础设施的技术架构。推理服务对时延极度敏感，智能对话要求首 Token 延迟低于 100 毫秒，这要求集群具备超低时延与高带宽的互联能力，以及毫秒级的动态调度能力。头部云厂商正加速推理引擎与调度系统的技术创新。华为在 2026 MWC 上海展期间联合湖北移动完成全国运营商首个 AI 推理加速方案现网测试，基于 OceanStor A800 存储与昇腾 A3 超节点架构，搭载 UCM 推理记忆数据管理技术，在长序列 AI 推理场景下实现首 Token 延迟优化 26%-93%。阿里云在 2026 世界人工智能大会上展出百炼推理平台，依托超节点算力集群兼容主流 AI 生态，已服务全球 500 万用户、120 万付费用户。

算力服务商业模式正经历从“卖硬件”到“卖服务”再到“卖价值”的三级跃迁。从裸金属 GPU 租赁（如按照机柜/月计费）、算力规模租赁（如按照 PFLOPS/年计费）、GPU 时租（如按照 GPU/小时计费），

演进到 Token 即服务（TaaS，按 Token 消耗量计费）、智能体即服务（AaaS）、结果即服务（RaaS）。这一跃迁在产业层面集中体现为三大动向。一是运营商全面入局。2026 年 5 月，中国电信、中国移动、中国联通相继推出 Token 套餐产品，中国电信最低月费 9.9 元含 1000 万 Token，中国移动最低 5 元体验包，Token 首次被纳入通信资费体系。二是独立 Token 服务商崛起，硅基流动定位为开放独立的 Token 供应平台，不研发基础大模型，依靠自研推理引擎与算力编排系统整合多元异构算力。三是上市公司加速卡位。润建股份、弘信电子、行云科技、超讯通信等 10 余家上市公司宣布探索 Token 工厂业务。

表 2 算力服务商业模式演进情况

演进阶段	商业模式	计费单位	核心特征	典型代表
第一阶段	裸资源租赁	机柜 / 月、PFLOPS / 年	售卖硬件资源，重资产交付	数据中心机柜出租、算力集群整包
第二阶段	算力即服务	GPU / 小时、算力时长	售卖算力服务，支持弹性调度	云厂商 GPU 实例、智算中心算力租赁
第三阶段	Token 即服务（TaaS）	Token 消耗量	售卖语义产出，按价值计费	Token 工厂、推理服务平台

二、智算基础设施政策导向

2025 年以来，国家密集出台政策推动 AI 与实体经济深度融合，智算基础设施的政策体系正从“重建设、轻运营”全面转向“以用促建、以 Token 效益定优劣”。“人工智能+”行动从顶层设计走向全面实施；“算电协同”首次写入《政府工作报告》并被纳入“十五五”规划纲要；“东数西算”工程持续深化，算力网被纳入国家“六张网”重点基础设施。与此同时，地方政策着力点正从建设规模转向服务效能，从补贴建算力转向采购 Token 服务。面向“十五五”，一个以 Token 为价值标尺的

智算基础设施政策框架正在加速构建。

（一）“人工智能+”行动：强化算力统筹与创新运营服务

“人工智能+”行动为智算基础设施的体系化、科学化部署提供了整体指引。2025 年 8 月，国务院印发《关于深入实施“人工智能+”行动的意见》，这是中国首份以“人工智能+”为主题的国务院文件，围绕算力提出“强化智能算力统筹”“优化国家智算资源布局”“加强智能算力互联互通和供需匹配”“创新智能算力基础设施运营模式”等系列要求。2026 年，“人工智能+”进一步上升为国家战略部署的核心内容。“十五五”规划纲要将“提升数智化发展水平”单独成篇，明确提出，要全面实施“人工智能+”行动，全方位赋能千行百业，促进生产方式深层次变革和生产力革命性跃迁。2026 年 7 月 30 日，中共中央政治局会议明确提出，要深入实施“人工智能+”行动，发展智能经济新形态。在推理需求成为算力需求主体的产业背景下，这意味着智算基础设施的发展逻辑正从以建设为中心的规模扩张，逐步转向以服务为中心的效能提升，重点在于支撑多少推理场景、交付多少高质量 Token 服务。

围绕算力资源的统筹配置与科学布局，国务院及各部委在“人工智能+”政策框架下明确了多项具体任务。国务院《关于深入实施“人工智能+”行动的意见》提出“支持人工智能芯片攻坚创新与使能软件生态培育，加快超大规模智算集群技术突破和工程落地”“充分发挥‘东数西算’国家枢纽作用，加大数、算、电、网等资源协同”。行业专项政策进一步细化落地路径，工信部等八部门《“人工智能+制造”专项行动实施意见》提出“有序推进高水平智算设施布局，加快建设全

国一体化算力网监测调度平台”，推动智能芯片软硬协同发展。交通运输部等七部门《关于“人工智能+交通运输”的实施意见》提出“统筹优化算力供给能力，动态整合企业算力，因地制宜强化算力保障”。国家药监局《关于“人工智能+药品监管”的实施意见》提出“统筹规划多级智能算力资源协同体系”，打造标准化、可扩展的智能算力底座。系列政策从国家枢纽协同、行业算力整合、多级算力体系建设等维度，为智算基础设施的科学布局与供需匹配提供了政策依据。

围绕智算基础设施的运营模式创新与服务能力提升，各部委在行业专项政策中作出针对性部署。国务院《关于深入实施“人工智能+”行动的意见》提出“创新智能算力基础设施运营模式”“培育人工智能应用服务商，发展‘模型即服务’、‘智能体即服务’等”。工信部《“人工智能+信息通信”创新发展实施意见（2026-2028 年）》围绕提升优化智算云服务作出部署，提出“推动平台即服务，打造以人工智能为核心的网络大模型和智能体开发工具链与自动化运维平台”。教育部等五部门《“人工智能+教育”行动计划》提出“建设国家教育智能算力服务平台（教育智联网），提供高质量算力支撑、数据服务、模型能力和智能体工具”。国家卫健委等五部门《关于促进和规范“人工智能+医疗卫生”应用发展的实施意见》提出“强化基础设施建设、优化人工智能算力算法”。系列行业专项政策从智算云服务、行业智能算力服务平台、模型即服务等维度，明确了智算基础设施在行业赋能中的运营模式与服务路径，推动算力资源从通用供给向行业适配的精细化运营延伸。

（二）全国一体化算力网：面向 Token 服务的算力跨域协同持续推进

国家层面将算力网定位为现代化基础设施体系的关键组成，全国一体化算力网建设正深刻重塑智算基础设施的运营逻辑与价值定位。2026 年 4 月，中共中央政治局会议明确提出加强算力网等规划建设；5 月，国务院常务会议将算力网纳入国家“六张网”重点基础设施。“十五五”规划纲要提出“加强全国一体化算力监测调度，提升算力接入和精准匹配能力”。全国一体化算力网的建设，正在将分散的智算基础设施从“单点孤岛”整合为全国统一的算力资源池，智算基础设施不再以单一集群为单位独立运营，而是作为全国算力网中的节点被统一发现、调度和按需使用，直接改变了智算基础设施的运营逻辑与价值定位。

工信部致力于推动智算基础设施互联互通。2026 年 2 月，工信部印发《关于开展国家算力互联互通节点建设工作的通知》，启动“1+M+N”国家算力互联互通节点体系建设，通过构建“统一标识、统一标准、统一规则”运行机制，实现不同区域、主体、架构的算力资源标准化互联和高效流动。2026 年 8 月，工信部公布首批 17 个国家算力互联互通区域节点建设名单，部分省市正式纳入全国算力资源统一调度体系。此外，工信部等七部门联合印发《促进平台经济大中小企业协同发展行动方案（2026-2028 年）》，提出“深入开展算力基础设施高质量发展和算力互联互通行动计划”“推进全国一体化算力监测调度服务平台建设”。工信部系列政策从标准制定、节点布局、平台

建设等维度，为智能算力跨域调度与 Token 服务的规模化发展提供了制度框架与基础设施支撑。

全国一体化算力网监测调度平台建设取得阶段性进展，Token 级调度正在成为智算基础设施运营能力升级的关键方向。据国家数据局介绍，20 余家算力调度主体已正式入驻全国一体化算力网监测调度试验验证平台，平台以 63 项指标绘制全国算力资源“动态地图”，初步实现对八大枢纽、十大集群及三大运营商等主体的监测调度。广州上线全国首个基于“词元（Token）”级调度的城市综合算力运行服务平台，以 Token 为统一计量基准构建计费体系，实现全栈异构算力的统一纳管与资源池化。江苏上线全国首个省市一体化算力调度监测平台，已监测全省 282 个数据中心、算力规模达 87.3EFLOPS，并与国家算力调度平台实现融合贯通。平台建设的持续深入，正在将智算基础设施的调度粒度从任务级细化至 Token 级。运营方需要以 Token 为最小单元进行资源感知、负载均衡与计费结算，这对智算基础设施的调度系统、计量体系和服务交付能力提出了全新要求。

（三）地方政策试点：突出智算基础设施运营服务效能

政策的激励标的正从算力建设规模转向 Token 运营效能，设施产出能力成为政策支持的直接依据。北京经济技术开发区 2026 年 8 月发布全市首个“词元经济”专项政策（“词元十条”），明确到 2030 年全面推动体系化、标准化、市场化的词元分发与消费基础设施建设；高标准推动万卡级词元工厂建设，并对开展推理效能跃升的企业按算力租赁费用的 30% 给予最高 2000 万元资金支持，激励依据不再是机柜

数量与峰值算力，而是单位能效下的词元产出效率。四川省发改委 2026 年 8 月印发《四川省推进算力网建设的若干政策措施（试行）》，支持建设词元（Token）工厂等算力公共服务设施，省级预算内基本建设投资按照项目核定总投资的 15% 给予支持。政策信号的转变，正引导存量智算资源加快转化为标准化、可计量的 Token 服务能力。

政策工具正从支持单点设施建设转向培育平台化运营体系，推动分散算力聚合为城市级 Token 服务能力。上海依托市级智能算力资源统筹调度平台，将算力券、模型券（Token 券）、语料券申报与全市算力资源统一接入、统一调度相衔接；2026 年 8 月上海市民间投资行动方案进一步提出推广发放三类券种，降低公共数据、算力、模型、语料等数字要素使用成本。广州海珠区“词元八条”覆盖词元“生产-流通-消费-增值”主要环节，支持企业开展异构算力词元化封装、多模态模型聚合接入、一站式结算等业务。深圳宝安区引导高水平算力服务商从“硬件设备交付”向“模型即服务（MaaS）”模式升级。政策的着力点正落在计量、调度、分发等运营环节，而非单纯的设施扩容。

试点示范成为智算设施网络化升级的重要路径，推动算力中心从单点部署向协同服务演进。2026 年 3 月，工信部发布《关于组织做好 2026 年度城域“毫秒用算”专项行动的通知》，明确首批在 31 个省（区、市）的 50 个地市开展试点，推动算力中心间毫秒互连（<1 毫秒）、重点场所毫秒接入（<1 毫秒）、终端应用毫秒可达（<10 毫秒），为低时延 Token 服务夯实城域网络底座。各试点城市围绕专项行动目标推进智算基础设施网络能力升级，浙江杭州、宁波、嘉兴三市入选。

其中，宁波移动联合其他运营商持续优化城域算间网络布局，完善鄞州、浙东、杭州湾机房等算力中心间互连架构，实现端到端高速通道互联，将算力中心从“孤岛”升级为“超级枢纽”；引入全光高速大容量无损传输（HIC-OTN）新技术，部署汇聚层星环波分、细颗粒波分（FG-OTN），2027 年汇聚覆盖目标达 100Gbps。试点将政策目标转化为可检验的设施能力标准，推动智算基础设施从单点部署向网络化协同升级，为低时延 Token 服务在制造、金融、医疗等行业的规模化应用夯实网络底座，也为同类城市提供了可复制的参照样板。

三、智算基础设施区域布局优化

推理算力正成为智算需求的主体，其“贴近用户、就近部署”的布局要求，正在重塑智算资源的空间配置逻辑。训练任务高度集中，推理任务广泛分布，二者对时延、成本、部署位置的要求存在差异。布局优化需以 Token 需求分布为核心依据，在国家算力枢纽与非枢纽区域之间形成差异化布局路径，并通过跨区域协同机制实现全国 Token 服务资源的最优配置。

（一）布局逻辑重构：训练与推理的空间分化倒逼资源配置范式转变

训练任务高度集中而推理任务广泛分布，智算基础设施区域布局需进一步优化。大模型训练任务主要由北京、上海、深圳等头部城市 AI 企业模型研发需求牵引，并依托西部枢纽的规模化集群承载，这些区域对算力密度和集群稳定性要求极高。推理任务则呈“星罗棋布”分布，实时交互、智能制造等场景广泛分布于人口密集、产业发达的

城市及边缘节点。国家数据局数据显示，2026 年 6 月中国日均 Token 调用量近 175 万亿，且调用热力随人口密度、产业活跃度和应用场景呈显著地理不均。空间分化的核心驱动力在于：训练追求“低成本、高密度”，可远离用户；推理追求“低时延、高并发”，需要贴近用户部署。两类需求难以在同一布局原则下充分兼顾。

传统“靠近电力与气候优越区域”的布局原则在推理时代的适用性面临局限。过去十年，智算基础设施选址主要遵循成本导向，西部枢纽凭借低廉电价与凉爽气候成为训练算力的主要承载地。这一逻辑在训练时代行之有效，但推理服务对时延高度敏感，端到端的首 Token 延迟由网络传输、Prefill 计算、Decode 生成、调度排队等多环节累加构成，即使枢纽间网络时延已优化至 20 毫秒以内，留给计算与调度的时延预算仍极为紧张。在高并发生产环境中，Prefill 计算排队和调度开销很容易突破剩余预算，使得西部枢纽在承载东部低时延推理需求时面临物理距离带来的时延约束，需从布局层面进行针对性设计。“十五五”规划纲要提出“加快‘枢纽-区域-边缘’算力设施协同布局，引导大型及以上算力设施在枢纽节点集聚”，已从顶层设计层面对这一差异化布局作出回应。

布局逻辑的重构正从三个方向推动全国智算基础设施的空间形态变革。一是地理布局上，训练算力相对集中、推理算力就近部署的分工日益清晰。训练任务宜依托枢纽节点规模化承载，推理节点则向东部及中部产业发达、人口密集区域集约化部署。二是架构布局上，推理算力加速向边缘和端侧下沉，数亿台 AI PC 和 AI 手机构成分布

式边缘算力池，端侧算力从“闲置资源”向“可用产能”转变，形成“云-边-端”三级协同的算力架构。三是产能形态上，西部枢纽从“算力基地”向“Token 产能基地”转型，绿电优势不再仅体现为降低训练成本，更指向支撑大规模、低成本 Token 的工业化生产，以绿色 Token 服务的形式面向全国输出。三个方向的变革相互叠加，正在重塑智算基础设施的空间形态与服务能力。

（二）算力枢纽节点：从“东数西算”到“训推分层协同”的分工深化

八大国家算力枢纽节点已形成智能算力的主要承载格局，枢纽内部的训练与推理分工正逐步显现。据国家数据局在 2026 中国国际大数据产业博览会上介绍，截至 2026 年 7 月底，八大枢纽和三个算电协同区已建成智算规模占全国比重超 85%。京津冀、长三角、粤港澳大湾区等东部枢纽仍是算力需求核心区，内蒙古、宁夏、甘肃、贵州等西部枢纽则是算力供给增长极。2024 年至 2025 年，各枢纽节点建设以训练集群为主。然而，随着推理需求持续增值，枢纽节点的功能定位正在从“训练为主”向“训推协同”方向调整。东部枢纽在现有训练集群基础上叠加部署高密度推理节点，面向本地金融、制造、互联网等产业提供低时延 Token 服务；西部枢纽依托绿电与气候优势，在持续承载训练任务的同时，探索向“Token 产能基地”转型，通过算电协同机制将绿电优势转化为低成本 Token 产能。

各枢纽节点正在形成差异化的发展路径，在智算基础设施体系中承担不同角色。京津冀枢纽以北京为核心，依托头部 AI 企业与科研

机构的模型研发优势，侧重“训练+高价值推理”功能；张家口、怀来等环京地区凭借毗邻北京、气候凉爽的区位条件，主要承载非实时推理与温冷数据存储，形成与北京核心区互补的分层布局。长三角枢纽以上海为龙头、苏州、嘉兴、芜湖等城市为支点，侧重训推一体平台化运营，上海“模塑申城”工程推动算力平台与行业场景对接，嘉兴探索区域性 Token 公共服务平台，芜湖作为“东数西算”十大集群之一重点承接长三角地区算力需求，形成从训练集群到边缘 Token 节点的完整链条。粤港澳大湾区枢纽以广州、深圳为核心、韶关为起步区，侧重服务智能制造、金融科技等产业的低时延推理需求，依托韶关集群的大规模算力供给与广深的低时延网络，构建“韶关训练+广深推理”的协同格局。成渝枢纽采用天府、重庆双集群布局，面向西部区域的行业智能化需求提供算力支撑。西部四大枢纽结合绿电与气候优势，侧重低成本 Token 产能的规模化输出，内蒙古和林格尔定位北方绿色算力中心，贵州贵安智算占比居全国前列，甘肃庆阳着力打造面向全球的 Token 工厂，宁夏中卫聚焦绿色智算高地。八大枢纽各有所重，推动国家算力布局从统一建设走向功能分化。

（三）非枢纽节点区域：三级梯度布局支撑 Token 服务广泛落地

对于非算力枢纽节点的广大区域，需根据城市能级与产业特征进行差异化定位，构建“核心城市-重点城市-活力城市”三级梯度的 Token 服务布局体系，实现与枢纽节点的错位互补。

核心城市（北京、上海、深圳）以超大规模训练集群与高密度推

理节点叠加部署，承担前沿模型研发与高价值 Token 服务供给的双重使命。这三座城市汇聚了全国绝大部分基础大模型研发力量与 AI 独角兽企业，是超大规模训练集群的核心承载地，也是 Token 调用热力最高的区域。北京已明确建设 Token 工厂、推广 Token 券；上海持续推进“算力浦江”专项行动，人工智能产业规模居全国前列；深圳构建城市级算力统筹调度平台，汇聚全市多元异构算力。核心城市通过“训练-推理”双能力协同，以前沿模型研发直接驱动高价值 Token 服务供给，实现从“模型创新”到“Token 交付”的最短路径闭环。

重点城市（成都、南京、杭州、宁波等）以行业模型推理与微调 Token 供给为主，兼顾区域训练需求，构成 Token 服务的区域支点。这些城市产业基础雄厚、科教资源丰富，是行业大模型与区域 AI 应用的聚集地。其核心功能是将通用大模型与本地行业场景深度结合，提供行业模型微调与推理 Token 服务。中国联通在宁波建成日产百亿级的 Token（词元）工厂，搭建自研 Token 网关，成为全省乃至全国较早落地 Token 服务的城市，上线 1 个多月已服务上千家企业、10 余个园区，其中近 50%用量来自制造业企业，跨境电商、外贸占比超 20%。重点城市“承上启下”，向上承接核心城市的模型能力输出，向下将 Token 服务转化为区域产业数字化的实际生产力。

活力城市（地市级及区县级）以边缘 Token 节点服务本地场景，构建“少训练、多推理、泛在边缘”的基层 Token 服务网络。地市级及区县级区域算力需求呈碎片化、场景化特征，如智慧城市视频解析、基层政务智能问答、文旅数字人接待、农业 AI 诊断等，对 Token 服

务提出低时延、数据不出域、低成本等要求。广东湛江联合云天励飞落地面向政务的“国模+国芯”人工智能词元工厂，以“百亿 Token 一分钱”为目标导向，自主研发 GPNPU 推理架构，通过对推理效率和能效比的深度优化，显著降低单位 Token 成本，推动 AI 深度渗透海洋、应急、教育、政务、自然资源等核心领域。活力城市的 Token 服务布局已从个别试点走向规模化铺开，以推理优先的差异化定位、灵活多样的投建模式和贴近场景的服务能力，就近满足本地推理需求，成为 Token 服务普惠化覆盖的关键一环。

（四）跨区域协同：强化 Token 供需资源优化配置

推理需求的潮汐性与地域性差异，使跨区域 Token 调度成为资源优化配置的必然选择。Token 调用热力与区域经济活跃度高度正相关。东部沿海地区在工作日白天形成 Token 需求高峰，西部地区受产业结构影响需求曲线存在差异。文娱类应用在节假日和晚间集中爆发，金融、政务等场景则在工作日形成稳定消耗。这种时空分布差异，为跨区域调度提供了“削峰填谷”的空间。当东部城市进入夜间需求低谷时，可将闲时 Token 资源调度给西部仍处业务高峰的城市，或分配给非实时推理任务，如批量数据处理、离线内容生成等。与此同时，东部地区算力供不应求、西部地区算力相对富余的结构性矛盾长期存在，仅靠本地化部署难以兼顾成本与体验，跨区域调度是兼顾二者的必经之路。

全国一体化算力调度体系为跨区域 Token 协同奠定了机制基础。国家数据局提出“全国算力资源一本账，算力调度一盘棋，算力整合

一张网”，全国一体化算力网监测调度平台已形成“一基座两能力”的架构。当前调度粒度仍以算力任务为主，难以满足 Token 级实时调度需求，但在资源感知、跨域纳管等方面积累了关键能力，为 Token 级协同的机制落地提供了基础支撑。各地算力调度平台也在探索跨区域资源对接，如湖北算力互联互通平台已接入武汉、宜昌、襄阳三大算力集群，安徽算力统筹调度平台集通算、智算、超算、量算“四算合一”，已与沪苏浙地区相关平台实现互联互通。

跨区域 Token 协同的规模化落地仍需在制度层面实现突破。当前面临的挑战有三个。一是质量保障机制缺失，不同区域集群的芯片代际、算力规模、网络条件差异显著，跨域场景下 Token 生成速率、首 Token 延迟等 SLA 指标难以统一承诺。二是成本分摊机制尚未建立，各区域电力成本、算力价格差异显著，缺乏公平透明的跨域结算规则与收益共享机制。三是跨域调度规则尚不完善，当多个区域同时发起 Token 调度请求时，优先级如何判定、调度顺序如何安排、跨域带宽如何分配，均缺乏系统性的规则设计。北京市《关于加快智能体引领发展的若干措施》提出“构建跨域协同算力供给体系”“推动小散算力‘零存整取’”，深圳、上海等地也在探索跨区域算力调度制度创新。只有同步推进跨域 Token 服务标准建设、算力交易结算机制完善与调度规则设计，才能推动全国 Token 服务资源的优化配置与价值共享。

四、智算基础设施技术创新突破

面向 Token 服务，智算基础设施技术体系的重心正从“峰值算力最大化”全面转向“每瓦 Token 产出最大化”。Token 服务低延迟、高吞

吐、低成本的核心诉求对智算基础设施技术栈的每一个层级都提出了新的要求。在芯片层面，存储带宽与内存容量取代峰值算力成为首要瓶颈，异构协同成为降低 Token 生成成本的核心路径；在软件层面，推理引擎与模型压缩技术决定 Token 的实际吞吐与质量，全栈协同优化填补标称性能与交付性能之间的鸿沟；在系统层面，推理服务器与集群架构从“通用计算节点”演变为“Token 生产单元”，超节点与训推一体成为主流部署形态；在网络层面，高速互联与 Token 级调度系统从“配角”升级为“主角”，在单集群极致性能的基础上进一步将能力扩展至跨集群协同，支撑毫秒级响应的工业化 Token 生产。

（一）芯片：异构协同降低 Token 生成成本

Token 服务的爆发对推理芯片设计提出了与训练不同的要求。训练任务追求峰值算力与大规模并行效率，而推理任务的核心诉求是低延迟、高吞吐和能效比，也就是用户不关心芯片的峰值算力是多少，只关心 Token 多快生成、成本多低。这一转变正在重塑芯片设计的技术优先级，存储带宽与内存容量取代峰值算力成为首要瓶颈，能效比（每瓦 Token 产出）取代单卡性能成为核心 KPI。产业界已初步形成共识，如英伟达 CEO 黄仁勋在 2026 年 GTC 大会上明确宣布，“每瓦 Token 产出效率”取代峰值算力成为数据中心的新度量标准。

异构协同成为推理芯片降低 Token 成本的核心路径。推理任务内部存在显著的计算特性差异。预填充（Prefill）阶段是计算密集型，负责将用户输入处理为 KV Cache，对算力密度要求高；解码（Decode）阶段是访存密集型，逐 Token 生成回复，对内存带宽和延迟极度敏感。

单一架构难以同时高效处理两类任务。英伟达 Rubin GPU 负责高并发预填充，Groq 3 LPU（语言处理单元）专攻低延迟 Token 解码，两者协同构成 GPU+LPU 异构推理架构。华为昇腾 950 同样采用“一芯双构”策略，950PR 主攻 Prefill，950DT 面向 Decode 优化。异构计算正从 GPU 统一处理走向最优芯片处理最优任务的专业化分工。

国产推理芯片在细分赛道加速形成差异化竞争力。推理芯片的竞争逻辑已从通用能力追赶转向特定场景突破。寒武纪聚焦推理赛道，其新一代旗舰云端 AI 芯片思元 690 已于 2026 年初实现量产，采用双 die 封装设计，配备 196GB HBM3 高带宽内存，实测互连带宽超 890Gbps，已完成 GLM、DeepSeek、Qwen、Kimi、MiniMax 五大国产主流大模型的推理适配。国产推理芯片在细分赛道形成了差异化突破，虽然单卡绝对性能仍有差距，但在特定精度、特定模型规模和特定能效比要求下已具备商业化替代能力。

（二）软件栈：智算软件体系加速重构升级

传统以训练为中心的智算软件栈侧重分布式并行训练（数据并行、模型并行、流水线并行等），其优化目标是缩短训练耗时。而面向 Token 服务的智算基础设施推理软件栈则需要围绕低延迟、高吞吐、低成本进行系统性重构。这种重构并非对训练软件栈的简单裁剪，而是从底层算子调度到上层模型服务的全链路变革，从“一次训练跑几天”到“毫秒级响应千万级并发”，对软件栈的响应速度、调度粒度和资源管理方式都提出了质的不同要求。

模型开发框架及工具链直接影响 Token 服务的部署效率与运行

性能。主流 AI 框架及其生态工具链负责将模型导出、转换、量化并部署，使其能够在不同芯片上运行。模型格式的标准化与跨平台兼容性直接影响后续推理效率。面向 Token 服务的大规模推理场景，需要缩短“模型发布到 Token 服务上线”的周期，华为昇腾 CANN 工具链提供从模型迁移、算子适配到图优化的全流程支持，百度昆仑芯 XPU 生态提供 PaddlePaddle 与 PyTorch 的无缝对接能力。与此同时，模型压缩技术正从可选优化变为必选能力。量化将权重从高精度压缩至低精度，使存储空间和内存带宽需求成倍下降；剪枝系统性地移除不重要的参数，减少模型冗余；知识蒸馏则将大模型知识迁移至更小模型中。三者协同可在精度损失可控的前提下大幅降低 Token 生成成本。英伟达新一代 Vera Rubin 平台在 NVFP4 精度下推理算力达 50 PFLOPS，混合专家模型的 Token 生成成本可降至 Blackwell 平台的十分之一。国产算力生态中，昆仑技术基于昇腾 AI 的 Ascend C 算子编程语言研发出软 FP8 解决方案，单台服务器即可流畅运行满血版 DeepSeek V3.1 模型，显存需求减半、端到端推理效率提升 32%。

软硬协同适配是推理软件栈能否充分发挥效能的基础保障。芯片峰值算力并不等于系统交付吞吐，同一硬件平台上，算子库、编译器、推理框架与模型的适配深度不同，实测 Token 吞吐可相差数倍。算子库的丰富程度决定了芯片的“可用算力”有多大比例能被实际调用；编译器的优化深度决定了计算图能否以最高效的方式映射到硬件指令；推理引擎的调度策略需要与芯片的内存层次、互联拓扑深度协同。寒武纪思元 690 完成五大国产主流大模型推理适配，从算子融合、内存

分配到调度策略进行端到端优化；华为昇腾通过 CANN 持续迭代缩小标称性能与交付性能之间的差距。面向 Token 服务的智算基础设施，亟需构建“芯片-算子-引擎-模型”全栈协同优化体系，以 Token 吞吐、首 Token 延迟、每瓦 Token 产出等实测指标为牵引，推动软件栈各环节的协同优化与持续迭代。

推理引擎的运行时调度能力是决定 Token 实际吞吐与响应时延的关键软件环节。传统训练调度以任务为粒度，追求大规模并行下的线性加速比；推理调度则需要以单个请求或 Token 为粒度，追求毫秒级响应下的资源利用最大化。因此，推理引擎的核心调度机制正在发生深刻变革，智算基础设施推理引擎的调度能力决定了芯片算力能否被有效转化为 Token 产出。连续批处理通过动态合并并发请求，消除传统静态批处理中等待慢请求的空转损耗；KV Cache 分页管理借鉴操作系统内存管理机制，消除显存碎片并支持更灵活的显存复用，大幅提升单卡并发处理能力；投机解码与多 Token 预测通过小模型“草稿”与大模型“验证”的协同机制，在相同算力消耗下实现数倍于传统自回归解码的 Token 生成速率。这些技术已在 vLLM、SGLang、TensorRT-LLM 等主流推理引擎中广泛落地，将实测吞吐较传统静态批处理模式提升数倍。2026 年，SGLang 为月之暗面 Kimi K3 提供发布日推理支持，通过混合注意力统一内存管理与内核逐级优化，将单卡解码速度从 44 提升至 113 token/s，叠加 DSpark 投机解码后进一步达到 423 token/s。

（三）集群：超节点与训推一体重构部署形态

单芯片能力的提升需要服务器与集群架构的系统性配合才能释放。面向 Token 服务，推理服务器正从通用计算节点演变为 Token 生产单元，集群架构从“训练优先”转向“训推一体”。

推理专用服务器与液冷技术成为 Token 工厂的标配。英伟达新一代 Vera Rubin NVL72 机架集成 72 颗 Rubin GPU，采用 100%全液冷设计并支持 45°C温水冷却，无线缆中板等模块化设计将计算托盘组装时间大幅缩短，支撑吉瓦级 AI 工厂快速部署；AMD Helios 机架集成 72 个液冷 MI450 GPU，单柜 FP4 峰值算力达 2.9 EFLOPS，2026 年下半年起率先在甲骨文云等平台规模化部署。国产阵营同步跟进，新华三 UniPoD S80000 超节点支持单机柜 32 卡至 128 卡高密度部署、最大可扩展至 16384 卡，新一代产品 PUE 可控制在 1.04 以下；浪潮计算机构建覆盖芯片冷板、机柜管路、CDU 与机房的液冷全链路能力，冷板式方案可支撑单芯片 1000W 以上散热，实现 $PUE \leq 1.20$ ，并支持存量机房灵活改造。

超节点架构成为集群形态重构的核心方向。超节点通过高速互联协议将数十至数百颗加速芯片整合为统一编址、低时延、高带宽的协同计算系统。2026 年被业界定义为“国产超节点元年”。华为 Atlas 950 SuperPoD 昇腾超节点单柜 64 卡单元可横向扩展至 8192 卡，支撑 50 万卡国产智算集群。中科曙光展出国内首套十万卡全栈自主“登峰 8000”AI 超集群。中兴通讯智算超节点方案以多芯开放协同设计，单机柜支持 128 个 GPU，可 Scale-up 至 1.6 万卡。中诚华隆 HL200 超

节点单机柜支持 64 卡互联，单节点最高 1024 卡纵向堆叠，横向可扩展至 10240 卡。

训推一体架构正在成为 Token 工厂的主流部署模式。传统智算中心往往将训练与推理算力割裂配置，导致资源利用率低下。训推一体平台通过统一调度，在训练任务间隙将算力切换给推理任务，或将推理闲时的算力用于模型微调。新华三 UniPoD S80000 系列定位于“万亿参数模型训推一体算力底座”。京算 Token 工厂通过异构算力池化调度，将资源利用率从行业普遍的 30%-40%提升至 80%以上。训推一体已成为 Token 工厂的核心部署形态，直接决定了算力资源的整体利用效率。

（四）互联：跨集群协同推进 Token 级调度

Token 服务的规模化运营要求智算基础设施突破单集群的物理边界，将分散在不同地理位置的算力资源整合为统一的 Token 产能网络。在这一架构下，网络互联解决 Token 如何在芯片、服务器、集群之间高效传输的数据流动问题，调度系统解决 Token 应该交给哪颗芯片、哪个集群去处理的任务分配问题。网络互联与调度系统的协同优化，正将智算基础设施从“单集群算力孤岛”升级为“全国 Token 产能一张网”，共同构成了智算基础设施实现 Token 工业化生产的运行底座。

高速互联网络是跨集群协同的物理基础，决定了 Token 在集群间、区域层面传输的效率。集群层面，超节点内部 Scale-Up 互联已在上一节阐述，本节的侧重点在于跨集群的互联协同，全国算力枢纽间 20 毫秒时延圈已基本实现，为跨地域的 Token 调度提供了网络前提。区

域层面，湖北“1ms 城市算网”加快落地，城市内中型及以上算力中心光层直连覆盖率达 100%，使同城多集群间的 Token 负载均衡成为可能。骨干网层面，呼和浩特国家级互联网骨干直联点网间互联带宽由 600G 扩容至 2300G。边缘层面，边缘 Token 节点依托云边端协同架构，通过算力感知路由、分布式推理网关等技术实现本地推理与跨域协同。网络性能直接决定了跨集群 Token 调度的可行性与效率，当调度系统将 Token 请求分发给远端集群时，网络的带宽和延迟决定了用户能否在可接受的时延内获得响应。

Token 级调度系统是跨集群协同的“大脑”，决定了 Token 请求如何在分散的算力资源间实时分发与负载均衡。传统训练调度以任务为粒度，一旦分配便长期占用资源；Token 服务则需要以单个请求甚至单次 Token 生成为粒度进行实时调度。所谓 Token 级调度，是指每一次生成 Token 的前向传播都是一个独立的调度决策点，调度器在每生成一个 Token 后，动态决定下一步用哪颗芯片、哪个集群继续处理，使 GPU 在毫秒级时间切片内轮转服务于多个不同模型的推理请求，始终处于满负荷运行状态。这意味着调度系统需在毫秒级完成路由决策、负载感知与故障迁移。在资源调度层面，异构算力统一池化、跨节点 GPU 资源池化、弹性伸缩将分散在不同集群的 GPU 资源抽象为统一资源池。在任务调度层面，结合拓扑感知、资源感知、负载感知、优先级感知的多维策略，在毫秒级完成路由决策与负载均衡。HAMI 商业化版本 Rise 已在央国企场景中实现 4 倍运行效率提升。

跨域协同创新正在推动 Token 生产从单集群走向分布式网络。运

营商已完成多项技术验证：中国联通在临港完成 AI 大模型 300 公里分布式协同训练验证，跨域等效算力达单集群 95% 以上；中国移动完成业界首次 800G 以太网智算协同训练现网试验，等效算力效率达 98% 以上；中国电信联合无问芯穹完成国内首例超 4000 公里大模型跨域混训，四集群协同性能提升 41%。跨集群异构 PD 分离架构（Prefill-RelayDecode-MainDecode 三级分离）进一步将推理任务拆分为多个阶段并分布在不同集群执行，通过多级解码集群的协同调度，提升大规模 Token 服务场景下的跨域部署能力。跨域协同的突破意味着 Token 生产不再受限于单一集群的物理规模，当本地集群算力不足或电价处于高峰时，可通过网络调度将 Token 请求分发至远端集群，实现跨地域的产能弹性扩展与成本优化。

五、智算基础设施产业生态迭代提速

Token 服务的爆发式增长正在从技术路径、商业模式到产业生态全面重塑智算基础设施的发展格局。面向推理时代的智算需求，基础设施的建设交付从工地建造走向工厂生产，供电、冷却、IT 设施平台等配套体系同步完成能级跃升；产业主体格局深度重构，央国企从算力投资方全面转向 Token 服务运营商，传统 IDC 服务商从资源租赁向 Token 产能交付转型，新兴 Token 服务商与 CDN 厂商以技术创新和边缘网络优势入局，三类力量各展所长、协同互补；产业协作模式亦走向协同创新。算电协同、分布式协同、算模协同、政产学研用协同四种模式相互嵌套、层层递进，共同推动智算基础设施从资源供给走向价值交付。产业生态的系统性迭代，正在为 Token 经济的规模化

落地提供坚实支撑。

（一）基建范式：智算基础设施建设运营的全链条升级

面向 Token 服务的智算基础设施爆发式增长，正在推动从机房建设、电力供给、冷却散热到 IT 设施平台的全链条能力升级。传统数据中心以通用计算为主，建设周期长、标准化程度低、功耗密度低；而面向 Token 服务的智算中心需要更高功率密度、更快交付速度、更低运营成本，催生了基础设施建设运营模式的系统性变革。

建设模式从工程项目向工业化产品演进。Token 服务的爆发式增长对智算基础设施的交付速度提出了前所未有的要求。传统数据中心建设周期通常为半年至一年以上，难以满足算力需求的快速扩张。阿里云推出面向 AIDC 研发的 CUBE 5.0 新一代架构，通过全模块化设计，供电、冷却、安防、智能化以及消防五大系统的整体模块化率从早年的 30%提升至 90%，将大型 AIDC 数据中心的交付工期缩短至 100 天，整体建设成本较上一代降低超过 10%。太初元碁在 2026 年 8 月发布预制化集装箱算力单元，工厂预制率超 90%，可在 24 小时内完成投运，交付效率提升 70%。维谛技术推出 Vertiv™ OneCore 全模组、全预制方案，部署速度较传统模式提升 50%，基建系统成本降低 20%。智算中心正从“工地建造”变为“工厂生产、现场组装”的工业化产品。

供电系统从传统配电走向算电协同的电力新基建。智算中心单机柜功率密度从传统的几千瓦跃升至数十千瓦乃至上百千瓦，对供电架构提出了革命性要求。特锐德发布全球首创的“算电岛”一体化供电方

案，以高压直入、中压直转、直流直供、预制集成、SST 赋能、算电协同的全新架构，将变电站建设周期从传统 12 至 18 个月压缩至 5 个月，并实现词元用电成本下降 30%。秦淮数据联合美团等打造行业首个 SST（固态变压器）商用项目，首创 10kV 角形接入模式，省去传统供电链路中多个冗余转换环节，实现 10kV 交流市电到 800V 直流一步式转换，供电效率达 98.5%，占地面积较传统方案减少 50% 以上。英伟达已公开提出自 2027 年起推动支持兆瓦级机柜的 800V 直流数据中心供电架构。供电系统正从“被动配电”升级为“主动赋能”的算电协同新基建。

冷却系统从风冷为主转向风液融合与全栈液冷。智算中心高密度部署带来的散热挑战，使冷却系统成为决定 Token 工厂能否稳定运行的关键环节。传统风冷已难以满足单机柜数十千瓦以上的散热需求，液冷技术加速规模化部署。华东地区风液同源预制冷却方案落地最大规模的智算中心项目，创新采用风冷、液冷共用一套分布式冷却系统，液冷应用占比高达 85%，全年综合 PUE 稳定控制在 1.2。海尔发布“风液同源”AI 智算全栈冷却方案及磁悬浮 CDU 一体机，为大模型、AIDC 园区提供兼容风冷与液冷的一站式低碳冷却解决方案。格力推出风液融合创新方案，采用双回路动态供冷架构，风冷与液冷比例可实现宽域智能调节，配置 N+1 热备份，毫秒级无缝工况切换。冷却系统正从“辅助配套”升级为决定智算中心 PUE 和运营成本的核心系统。

IT 设施平台从单点管理走向全栈运营。面向 Token 服务的智算基础设施包含异构芯片、高速网络、分布式存储等多层次复杂系统，

传统单点管理工具已无法满足运营需求。华为发布 MindOps 智算运维解决方案，保障 AI 集群高可用和业务高长稳运行。中国移动研究院发布“芯合”2.5 平台，实现 9 家 AI 芯片的统一纳管与灵活编排，支持整卡、分卡、配额等细粒度调度能力。“联通星罗”先进算力调度平台已完成昇腾、昆仑芯、沐曦等 30 余款主流国产算力芯片全面适配，是业内率先实现四地远距离跨域异构混训的平台，千亿参数模型异构混训效率可达 85%以上。达卯科技推出算电协同 2.0 平台，以“融合推演智能+统筹博弈智能”双引擎架构，打通从数据感知到价值交付的全链路能力闭环。IT 设施平台正从“设备管理”升级为“产能运营”的核心中枢，支撑智算基础设施从资源交付走向价值交付。

（二）产业主体：三类力量引领智算基础设施角色转型

面向 Token 服务的智算基础设施爆发式增长，正在推动产业主体格局的深度重构。三类主体基于各自禀赋优势，在 Token 服务生态中形成了清晰的分工与互补格局：央国企凭借资本实力与全国性网络优势，从算力投资方全面转向 Token 服务运营商；传统 IDC 服务商依托数据中心资源与运维经验，从资源租赁向 Token 产能交付转型；新兴 Token 服务商以技术创新与敏捷迭代能力，快速构建起 Token 生产与流通的专业化服务能力。三类主体共同构成了从算力生产到 Token 交付的完整产业闭环。

央国企从算力投资方全面转向 Token 服务运营商。央国企凭借资本实力、基础设施资源和全国性网络优势，正在从投资建设智算中心的角色定位，升级为面向 Token 服务智算基础设施的核心运营力量。

集团层面：中国电信 2026 年 4 月启动国内首个以“Token 工厂”命名的 174.38 亿元集采项目，采购标的已从硬件设备变为 Token 生成能力服务；中国移动 2026 年 5 月发布 Token 运营生态体系，宣布未来 3 至 5 年投入百亿级 Token 生态资源、建设千亿级算力基础设施。地方层面：贵州移动率先将智算中心升级为 Token 工厂，完成算力资源池化、平台统一调度，截至 2026 年 7 月，贵州移动“词元工厂”日调用量已突破百亿，较年初增长 2 至 3 倍；云南移动打造 AI 算力服务平台 Token 工厂，搭建集 Token 规模化生产、智能调度、精准分发、全域运营于一体的全链路供给体系，通过异构算力统筹与潮汐智能调度，集群算力利用率从不足 30%提升至 90%以上。

传统 IDC 服务商加速向 Token 工厂转型。万国数据、世纪互联、秦淮数据、光环新网等传统数据中心服务商，正从“卖机柜、卖带宽”的资源租赁模式向“卖 Token 产能”的服务交付模式全面转型，其战略定位、业务模式与网络架构均发生系统性重构。战略定位层面，IDC 服务商正从基础设施提供者向 AI 算力服务运营者跃迁。秦淮数据明确提出从 IDC 向“Token 工厂”转型，以连接能源与算力的核心枢纽为定位，成为“东数西算”工程中投运规模大、运营效率高的算力中心企业之一。润泽科技 AIDC 算力业务 2026 年上半年同比增长 126.24%，智算中心正由单纯硬件机房升级为“Token 工厂”智能生产载体。网络与交付层面，IDC 服务商正通过算力网络升级将数据中心从“孤岛”改造为 Token 服务节点。世纪互联提出“超互联”体系，在宁波构建高性能算力池和大模型 Token 服务平台并部署 100G 超互联网络。世纪互

联同时推进绿色 Token 工厂建设，在江苏如东打造华东首个绿电直连与冷能直供一体化 Token 工厂项目，将 LNG 冷能智控技术融入算力基建。光环新网在京津冀等地区部署高性能算力硬件，算力业务规模已超 4000P，年合同额超 1 亿元，并在全国规划机柜超 23 万个。

新兴 Token 服务商从技术突破走向规模化运营，CDN 厂商携边缘网络优势入局。一批专注 Token 生产与流通的创业公司快速崛起。硅基流动定位为独立的 Token 供应平台，依靠自研推理引擎与算力编排系统整合多元异构算力，日均 Token 调用量达数万亿，2026 年 6 月完成超 20 亿元 B 轮融资，并向港交所递交上市申请。趋境科技半年内累计融资超 10 亿元，平均单台算力的 Token 生产效率提升 3 倍以上。清程极智构建了覆盖“Token 生产-Token 流通-Token 落地应用”的全链路技术体系。无问芯穹 MaaS 平台日均 Token 调用量 7 个月增长 40 倍。博伦智汇专注于异构智算平台与梯度 Token 工厂技术服务，自研 TOLD 技术栈已在万卡级别的英伟达与国产 GPU 异构集群上完成真实验证。与此同时，CDN 厂商正成为 Token 服务生态中不可忽视的新兴力量。网宿科技已在全球部署超 3000 个边缘计算节点，2026 年 7 月在常州落地首个城市级绿色词元工厂，综合试点绿电直连、800 伏高压直流、固态变压器、相变液冷等技术，项目建成后可年产 Token 60 万亿个。CDN 厂商的入局，标志着 Token 服务正从集中式智算中心向分布式边缘节点延伸，Token 的分发逻辑从“集中生产、远距离传输”走向“边缘生产、就近交付”。

（三）生态协作：三类模式引领智算基础设施协同创新

面向 Token 服务的智算基础设施建设运营涉及算力、电力、网络、模型、数据、场景等多个环节，单一主体难以独立完成全链条交付。当前，以智算中心建设运营方为核心的各主体积极推进与上下游主体的深度协作，探索形成三种主要的产业协作模式，共同推动 Token 工厂的高效建设与可持续运营。算电协同解决 Token 生产的成本基础，分布式协同解决 Token 产能的规模扩展，算模协同解决 Token 生产的质量效率，政产学研用整体协作解决 Token 服务的价值落地。

推进算电协同，智算基础设施与电力能源的深度绑定。面向 Token 服务的智算基础设施本质是将电力转化为 Token 产出，电力的成本与稳定性直接决定了 Token 生产的边际成本与连续性。在产业实践层面，商汤大装置在 WAIC 2026 发布算电协同 Agent，率先提出 TPW 指标以单位电力成本创造 Token 产出衡量 AI 生产效率；该 Agent 已在商汤临港 AIDC 全面落地，实现单位电力成本 Token 产出提升 80%、平均电费单价低于同地区 IDC 10%、算力负荷预测准确率达 96%，全年预估可实现碳减排 24000 吨/每万 P。常州签约首个城市级绿色词元工厂，综合试点绿电直连、800 伏高压直流、固态变压器、相变液冷、物理 AI 能源大模型等技术，项目建成后可年产 Token 60 万亿个，算力资源利用率从行业普遍的 30%-40%提升至 65%以上，等效 Token 产出增加 62.5%。算电协同正从政策倡导走向工程落地，成为 Token 工厂选址与建设的基础性协作模式。

推进分布式协同，以 Token 聚合平台实现分布式虚拟智算集群的

集约运营。面向 Token 服务的智算需求呈现“大规模集中训练+碎片化推理部署”的双重特征。分布式协同模式的核心是通过 Token 聚合平台构建分布式虚拟智算集群，将分散在不同地理位置的算力资源聚合为统一的 Token 产能池，实现算力资源的集约化利用与弹性供给。工信部在《关于开展普惠算力赋能中小企业发展专项行动的通知》中明确提出探索“算力银行”“算力超市”等创新业务，支持中小企业存入闲置算力资源，为分布式协同提供了政策依据。省级平台层面，湖北算力互联互通平台已接入武汉、宜昌、襄阳三大算力集群及三大运营商算力资源，实现可监测算力超 7000P、可调度算力超 2000P，平台将各规格底层算力封装为大模型 Token API 服务。城市级平台层面：“人民数据·城市词元工厂”平台将闲置算力封装成 Token 服务供企业按需取用；无锡城市智算云平台上线城市级“Token 超市”，汇聚智算资源超 13000PFLOPS，依据输入输出 Token 实际用量精准核算服务费用。第三方平台层面：共绩科技组建“电网式算力调度平台”，通过智能算法将分散的碎片化闲置算力整合为统一算力池，已接入 70 余万台设备（含个人电脑、网吧设备等）、50 余个智算中心，服务超 12000 家客户及团队，帮助客户降低算力成本 50%以上。这些 Token 聚合平台的上游连接各类分散的算力基础设施，下游辐射 AI 应用与智能体等终端场景，嵌合了 Token 工厂产业链中“算力调度与 Token 分发”的核心环节，正在将智算基础设施从“集中式大工厂”扩展为“无处不在的 Token 产能网络”。

推进算模协同，强化智算基础设施与模型算法的垂直整合。智算

基础设施推出 Token 服务的质量与成本，主要取决于算力与模型之间的适配深度。智算中心运营方正加速与模型厂商建立深度协作关系，通过联合调优实现 Token 生产效率的最大化。软通动力旗下北京壹号词元工厂与某头部大模型厂商签署智算服务协议，为其提供基于 Token 工厂的推理服务，涵盖大模型推理加速、高性能算力集群适配及行业 AI 应用落地等场景，共同推动从模型到 Token 服务的产业闭环。双方从算子适配、分布式通信到调度策略进行端到端调优，提升芯片在真实推理负载下的效率。优刻得发布词元工厂解决方案，构建覆盖“Token 生产-分发-调用-治理-运营”的完整基础设施，在底层与海光信息 DCU 芯片 BW 系列达成深度合作，通过对智谱 GLM-5.2、DeepSeek-V4 等主流大模型系统级适配与工程优化。摩尔线程与部分模型企业的合作已前移至研发阶段的联合共研，力争在模型发布之日即完成调优。商汤大装置联合近 20 家国产芯片及基础设施厂商建设 Token 运营中心，两年内规划 5 个万卡以上国产算力集群，将智算集群和模型服务纳入统一运营体系。智算中心正从“裸算力供给”转向“芯片-模型-场景联合优化”的深度协作模式，Token 服务的竞争力不仅取决于有多少算力，更取决于算力与模型的适配有多深、与场景的贴合有多紧。

推进政产学研用协作，突出智算基础设施与地方政府、行业客户的生态联合共建。智算基础设施要真正赋能实体经济，必须与行业场景深度结合。以智算中心运营方为核心，联合地方政府、行业龙头、科研机构、应用开发商等组建创新联合体，正在成为推动 Token 服务

落地的重要组织形态。常州绿色词元工厂由常州市数据局、常州高新区、国家电投电投综能、山铁数科、阿里云智能集团、网宿科技、深度动力七方合作，构建覆盖“绿电供给-算力建设-技术平台-运营服务”的全链条协同格局。中国移动发布“Token 运营生态体系”，联合腾讯、阿里、华为、中兴、科大讯飞等生态伙伴共同构建 Token 服务生态。政产学研用联合体通过整合多方资源，将智算基础设施的 Token 生产能力与行业场景需求深度绑定，推动 Token 服务从“通用能力供给”走向“行业 Know-how 嵌入”，实现智算基础设施赋能价值的最大化释放。

六、展望与发展建议

当前，面向 Token 服务的智算基础设施正处于范式重构的关键窗口期，但也应看到，面向 Token 服务的产业生态仍不成熟，智算基础设施的规划建设与服务效能之间仍存在结构性不匹配。建议通过系统性政策供给、区域布局优化、关键技术攻坚与产业生态升级，推动智算基础设施从“算力工厂”向“Token 工厂”高质量转型，为智算基础设施适应 Token 经济提供支撑。

（一）完善政策引导：强化规划协同与服务效能导向

优化智算基础设施规划布局，强化训练与推理的分类引导。结合前期各地智算中心建设运营的实践经验，建议在新建和改扩建项目中加强对训练型与推理型设施的分类引导。训练型智算中心侧重支撑大模型预训练等场景，可依托国家算力枢纽节点部署；推理型智算中心侧重面向实时交互、行业应用等场景提供 Token 服务，宜贴近用户侧集约部署。对于已建成的智算中心，可探索结合区域 Token 需求热力

开展训推功能适配的评估工作。通过分类引导，逐步构建“训练枢纽+推理节点+边缘 Token 端”的分层布局体系，推动算力资源配置效率的整体提升。

完善 Token 服务的政策工具组合，推动智算基础设施从建设导向转向效能导向。结合前期北京、上海、广州等地 Token 券、算力券的试点经验，建议在更大范围内探索将 Token 服务效能纳入算力基础设施支持政策的考量体系。政策着力点可从支持单点设施建设逐步向培育平台化运营体系延伸，推动分散算力经由平台汇聚转化为标准化 Token 服务能力。在政府购买算力服务方面，可前瞻研究将采购范围从基础算力扩展至 Token 服务，综合考量 Token 吞吐率、首 Token 延迟、每瓦 Token 产出等效能指标，以需求侧政策引导 Token 服务市场规范发展。

（二）优化区域布局：推动算力资源分级配置与跨域协同

推动算力枢纽节点从“训练为主”向“训推协同”功能演进。随着推理需求占比持续提升，算力枢纽节点的功能定位也需相应调整。东部枢纽在现有训练集群基础上可进一步叠加部署高密度推理节点，面向本地金融、制造、互联网等产业提供低时延 Token 服务；西部枢纽在承接训练任务的同时，可结合算电协同政策探索将绿电优势转化为 Token 产能优势。建议在枢纽节点规划中前瞻考虑训推一体的技术架构和部署模式，避免训练集群与推理算力割裂配置导致的资源利用不充分问题。枢纽节点之间的竞争维度正在从算力规模转向 Token 产出

效率，可适时引导这一转变，推动枢纽节点从“比算力规模”走向“比 Token 效能”。

支持非枢纽节点区域因地制宜布局 Token 服务设施。对于非算力枢纽节点的广大区域，可根据城市发展水平和产业基础差异化布局 Token 服务能力。重点城市可依托现有产业与科教资源，围绕优势行业（如智能制造、金融科技、智慧医疗等）部署面向行业模型推理和微调的 Token 服务设施，形成区域性的 Token 服务支点。活力城市可结合本地智慧城市、政务服务、文旅、教育等场景，部署 50-300P 的中小型推理节点或词元工厂，以较低成本就近满足本地推理需求。建议在活力城市层面，适度鼓励盘活存量算力资源，可通过 Token 聚合平台将分散的算力纳入统一调度，避免重复建设。

探索跨区域 Token 协同机制，推动 Token 资源优化配置。跨区域 Token 协同是破解东西部算力供需错配的重要方向，其实质是将分散在不同地理位置的算力资源通过调度机制整合为统一的 Token 产能池。建议在现有全国一体化算力网监测调度平台的基础上，逐步探索跨区域 Token 调度的机制设计，包括但不限于：跨域 Token 服务 SLA 标准、算力交易与结算规则、以及多区域并发调度时的优先级判定与带宽分配规则。鼓励有条件的地区开展跨区域 Token 调度试点，在实践中逐步完善相关机制设计。

（三）夯实技术创新：强化面向 Token 服务的软硬件基础

加强推理芯片与异构计算的技术攻关与生态建设。面向 Token 服

务对低延迟、高吞吐和能效比的要求，建议在国产推理芯片领域加大对存储带宽、内存容量等关键指标的攻关力度。异构协同是提升 Token 生产效率的重要方向，可结合 Prefill 与 Decode 阶段的计算特性差异，探索不同架构芯片的协同部署方案。在生态建设方面，建议加大算子库、编译器、开发工具链等基础软件投入，推动芯片从“可用”向“好用”演进，持续提升芯片在真实推理负载下的 Token 吞吐能力。

推进推理软件栈的全链路优化与软硬协同适配。从模型开发框架到推理引擎运行时的全链路优化，是释放芯片 Token 生产效率的关键环节。连续批处理、KV Cache 分页管理、投机解码等技术已在主流推理引擎中广泛应用，建议进一步推动其在国产算力平台上的适配与深度优化。算子库、编译器、推理框架与模型的适配深度对实测 Token 吞吐影响显著，可引导产业界构建“芯片-算子-引擎-模型”全栈协同优化体系，持续缩小标称性能与交付性能之间的差距。

推动服务器集群架构创新与绿色技术应用。建议面向 Token 服务场景推动服务器与集群架构的系统性创新。在集群架构层面，可围绕超节点高速互联、训推一体统一调度等方向开展技术验证与标准研究，探索在不显著增加硬件投入的前提下提升资源利用率的可行路径。在绿色低碳方面，建议结合算电协同政策，在新建及改扩建智算中心中探索绿电直连与 Token 生产的联动机制，根据机房条件适度引入风液融合、全栈液冷等散热方案，持续降低智算中心运营成本。

推进网络互联与 Token 级调度系统的协同优化。建议将跨集群互联能力纳入智算基础设施建设标准体系，推动 RDMA 等高速互联技

术的规模化部署，为跨地域 Token 调度提供网络基础。在调度系统层面，建议引导业界从传统任务级调度向 Token 级调度升级，通过异构算力统一池化、GPU 分片、弹性伸缩等技术手段，提升 Token 请求在分散算力资源间的实时分发与负载均衡能力，逐步形成可规模部署的 Token 级调度解决方案。

（四）培育产业生态：构建多元主体协同的 Token 服务体系

推动智算基建全链条工业化升级，夯实 Token 规模化生产底座。推广模块化、预制化智算中心建设模式，引导新建项目采用全系统模块化设计，缩短交付周期、降低建设成本，适配 Token 产能快速扩张需求。深化算电协同技术落地，推动高压直供、直流供电、固态变压器等方案规模化应用，将绿电直供、源网荷储一体化纳入智算中心建设标准体系，从供电侧降低 Token 生产边际成本。推动液冷与智能温控系统普及，支持高密度智算集群部署风液融合、全液冷方案，以 PUE 优化、单位算力能耗管控为抓手，持续提升 Token 生产能效水平。升级全栈智能化运营能力，引导智算中心部署统一算力编排与智能运维平台，实现异构芯片统一纳管、细粒度 Token 调度、故障自动自愈，支撑 Token 工厂稳定高效运行。

培育多元差异化产业主体，形成错位互补的 Token 服务格局。支持央国企发挥基础支撑作用，鼓励基础电信运营商、地方国资平台依托网络与资源优势，布局规模化 Token 工厂与公共服务平台，打造普惠性 Token 服务供给体系。推动传统 IDC 服务商向 Token 服务商转

型，引导数据中心运营商重构业务模式，从机柜租赁、带宽销售转向 Token 产能交付，支持依托存量机房资源改造升级 Token 服务能力。扶持新兴 Token 服务商创新发展，鼓励技术型企业聚焦推理引擎优化、算力调度编排等核心环节，以轻资产模式打造专业化 Token 生产与流通服务能力。引导 CDN 厂商布局边缘 Token 产能，发挥边缘节点广覆盖的网络优势，建设分布式边缘 Token 工厂，推动 Token 服务向“边缘生产、就近交付”延伸。

深化多维度产业协同模式，构建开放共赢的产业生态体系。深化算电协同模式，推动智算企业与电力企业建立深度合作机制，探索算力-电力联合调度与联合定价，推广单位电力 Token 产出（TPW）考核指标，实现电力成本与 Token 产能的联动优化。推广分布式协同模式，支持“算力银行”“Token 超市”等创新业态发展，鼓励城市级算力聚合平台建设，将分散算力资源整合为统一 Token 产能池，提升全社会算力资源利用效率。推动算模深度协同模式，引导智算运营方与模型厂商建立联合调优机制，将模型适配优化前置至研发阶段，实现模型发布日即完成生产级适配，最大化释放硬件算力的 Token 产出效率。完善政产学研用协同模式，支持地方政府、运营企业、科研机构、行业客户共建 Token 服务创新联合体，打通技术研发、产能建设、场景落地全链条，凝聚产业发展合力。

中国信息通信研究院 产业与规划研究所

地址：北京市海淀区花园北路 52 号

邮编：100191

电话：010-68033076

传真：010-68033959

网址：www.caict.ac.cn

