

算力租赁：高景气度下的趋势演绎20260901

2026算力租赁行业：推理需求驱动爆发，模式向Token 分成转型

摘要

- AI 推理需求爆发：2026年7月周度Token 调用量达50.7T，较2025年增长10倍；国内日均调用量两年增长超千倍， Agent workflow驱动算力从训练转向推理主导。
- 供给侧结构性分化：高端算力缺口超35%，H100缺口达43万张， B200 价格维持5.75美元高位；传统数据中心上架率不足60%，中低端算力出现结构性饱和。
- 市场规模与分布：预计2026年中国算力租赁市场规模超2,600亿元；东部地区算力设施占比55.9%居首，互联网行业贡献62%的核心需求。
- 商业模式转型：行业正从占比95%的裸金属租赁向“保底+Token 分成”模式演进，后者利润率预计为传统模式的两倍，旨在对冲硬件折旧并分享下游红利。
- 高杠杆融资特性：国内以设备融资租赁为主，海外倾向利用下游长约合同融资；金融创新正推动算力收益权质押、智算中心REITs 及算力残值保险落地。
- 核心标的观察：北美关注CoreWeave、LambdaLabs 及转型矿工 Iridian；A 股关注协创、利通、红警、高乐、软通动力等具备转型潜力的标的。

Q&A

当前算力租赁行业的整体景气度、市场阶段以及核心驱动力是什么？

算力租赁行业目前处于由供需共振驱动的高景气度阶段，并已进入订单落地与业绩兑现的周期。从需求端看，Token 调用量、AI Agent、多模态技术及行业应用的加速落地，持续拉动了推理算力的需求。同时，行业正处在技术升级和商业模式转型的关键时期，从早期的设备与卡时出租，向集群优化、异构调度及Token 服务等模式衍生。这种转型是行业穿越硬件迭代和算力价格波动周期，实现持续增长的关键。

从需求端来看，AI推理需求的增长趋势和主要驱动因素有哪些具体表现呢？

AI推理需求呈现出迅猛增长态势。根据OpenRouter平台的数据，模型周度Token调用量已从2025年8月约3-5T的区间，经过一年的爬升，到2026年7月27日当周达到了50.7T，实现约10倍量级的增长，且自2026年4月起增长曲线变得尤为陡峭。从国内数据看，根据国家税务总局的统计，我国日均词元调用量从2024年初的1,000亿，激增至2026年3月的1.4万亿，两年间增长超过千倍。这一增长主要由Agent工作流或代码任务主导的Token消耗驱动，推动AI应用从单轮交互向高Token密度的复合执行场景升级。随着泛办公、企业级RAG+ERP以及AI for Science等新场景的持续落地，需求端仍处在爆发初期。以Agent为核心的复合推理工作流大幅提升了Token密度和消耗连续性，驱动AI算力需求从训练驱动转向以推理为主导的韧性新周期。

当前算力市场的供给侧呈现出怎样的结构性特征？高端与中低端算力的供需状况有何差异？

供给侧呈现出明显的结构性分化：高端算力持续紧缺，而中低端及传统通用算力则出现结构性饱和。根据中国信通院数据，2026年第一季度，AI算力需求同比增长417%，而供给端增速仅为128%，供需缺口持续拉大。高端算力的缺口尤为突出，2026年第一季度缺口超过35%，其中H100芯片的市场缺口达到43万张。B200芯片价格在经历回调后，目前仍处于约5.75美元的高位，进一步表明高端算力的紧缺状况。相比之下，部分传统数据中心的上架率不足60%，平均出租率回落至40%以下，个别甚至跌至28%。不过，部分经典的或超期服役的GPU（如A100）仍具备市场需求价值。

造成前高端算力供给持续紧张的主要原因是什么？

高端算力供给持续紧张主要源于三个方面。首先是地缘管制，对高端及特供算力芯片、制造设备和云服务的出口限制，显著削弱了算力的可获得性。其次是成本上涨，半导体全产业链从上游代工到下游连接器普遍进入涨价通道，直接推高了计算节点的整机成本。根据芯查查数据，2026年1月至7月，半导体产业链普遍涨价5%至80%，这使得许多小型企业无力自行采购，转而寻求更具性价比的租赁方式。最后是芯片的代际切换影响，先进封装的良率爬坡与产能释放延迟，延缓了集群的交付节奏，导致高端算力在国内市场依旧非常紧缺。

从市场整体规模和地理分布来看，中国智算中心的发展现状和上架率情况如何？

智算中心规模呈现跨越式增长。根据工信部2026年7月的最新数据，全国算力设施整体上架率达到71.4%，反映出大模型训练和行业应用落地对算力资源的强劲支撑。从地理分布来看，算力资源主要集中在东部和西部地区。各区域算力设施规模占比分别为：东部55.9%、中部10.6%、西部32.6%、东北0.9%。东部地

区凭借其产业基础和AI应用需求优势，承接了大量算力部署；而西部地区则利用其资源优势，成为算力转移的主要承接地。

2026年中国算力租赁市场的预期如何？从行业需求结构来看，当前及未来的主要驱动力是什么？

根据中国信通院的口径，预计2026年中国算力租赁市场规模将超过2,600亿元。第三方机构如科智咨询的数据也印证了这一规模预测，显示行业整体呈现高增长态势。从行业需求场景来看，当前互联网行业以62%的占比成为绝对核心的需求方，其次是政府(14%)、金融(10%)、医疗(6%)，以及制造和教育(各占4%)。这表明目前的需求主要来自大型科技公司，整个市场仍处于爆发初期。未来，随着AI技术向金融、医疗、政府、交通、教育、工业制造等更多垂直行业渗透，整体需求预计将实现更大幅度的增长。

当前算力租赁行业的商业模式正在发生怎样的演进？不同模式的特点、优劣势及市场渗透情况如何？

当前算力租赁行业正处于从传统的裸金属租赁向更高级服务模式转型升级的关键起点。这一演进的核心在于提升服务的技术壁垒和运营效率，摆脱单纯的重资产和一次性买卖模式。算力租赁本质上是金融渠道与技术的复合商业模式。低层模式主要通过融资租赁为客户承担设备折旧和技术淘汰风险，赚取风险溢价；中层模式则利用获取高端算力(拉卡)的能力和供应链优势，赚取渠道差价。目前市场主流的裸金属租赁主要集中在这两个层面。高层模式则通过算力调度和池化技术提升集群利用率，赚取运营效率的收益。行业的服务模式正从以固定租金为主的裸金属租赁，逐步转向以Token分成或Token工厂一体化服务为代表的新型模式。

1. 裸金属租赁：这是当前最主流的模式，市场占比超过95%。其优点是现金流确定性高呈可预测，能在行业高景气度时享受涨价红利，缺点是收益天花板固定，需承担较大的资产折旧压力，在市场价格下行或剧烈波动时，面临客户违约和成本无法覆盖的风险。基于标准产值测算，该模式的毛利率大约在10%以上，但若按实际产值计算会高很多。
2. Token分成：这被视为行业发展的必然趋势，自2026年三四月份开始被提出，并于同年8月在部分头部厂商中实现了小范围落地。该模式下，算力租赁公司与模型厂商共享Token收益，相当于双方进行了一次期权交换。算力租赁公司借此摆脱纯粹的裸金属租赁风险，分享下游AI产业的增长红利，增强业务的可持续性。模型厂商则通过分享部分Token收益，锁定了未来算力价格上涨的风险。若保底收入能覆盖算力租赁的基本成本，Token分成模式的利润率预计至少是裸金属租赁模式的两倍。然而，该模式对算力租赁厂商的技术能力要求更高，涉及集群调试、客户模型API接入以及精确的分成计算等。

3. **Token 工厂**：目前这一概念尚处于非常早期的阶段，且市场定义较为宽泛。一些软件和应用公司提出的“Token 工厂”甚至包含AI应用、整体解决方案乃至机器人服务。而从算力租赁厂商的角度来看，纯粹的Token工厂模式是在Token分成模式成熟之后才会逐步尝试的更高级形态。

算力租赁厂商的业务模式从Token分成向Token工厂演进，对厂商自身提出了哪些新的要求？其优缺点分别是什么？目前行业主要处于哪个发展阶段？

Token工厂模式可以视为Token分成模式的进一步升级，它对算力租赁厂商在技术层面提出了更高的要求，包括需要进行集群化、云原生以及推理加速等改造。该模式的优点在于厂商能够全面享受Token带来的收益，理论上其净利率会高于Token分成模式。其缺点在于，一方面需要厂商自己解决客户获取问题，另一方面在底层技术调用和资源调度方面存在较高的技术壁垒。同时，公司也必须承担由利用率波动等因素带来的经营风险。从产业发展阶段来看，目前行业探索的重点仍然是Token分成模式。算力租赁公司普遍有推动业务转型升级的诉求，若能顺利完成向Token分成乃至Token工厂模式的转变，有望为其自身估值带来新一轮的提升。

从合同模式演进的角度看，算力租赁行业的合同模式经历了哪些发展阶段？当前主流的合同形式是怎样的？

算力租赁行业的合同模式演进大致可分为三个阶段，第一阶段以资源为核心，主要采用按整机或算力卡数量出租的单向风险模式。第二阶段为了规避价格波动和未来不确定性的风险，逐步发展为签订长期合约的模式，这也是当前市场的主流做法，尤其在海外，长期合约已成为重要的融资标准。目前，一种常见的合同形式是“3+2”模式，即前三年以锁定的长单价格覆盖GPU的折旧成本，后两年则根据当时的市场算力价格调整租金，这种方式相对灵活。第三阶段是算力租赁公司追求更高收益的阶段，倾向于采用“保底+Token分成”的模式，旨在与下游客户共同分享Token收益，这与服务模式的升级趋势相契合。

算力租赁行业高杠杆的特性体现在哪些方面？国内外市场在融资模式上有何差异？

算力租赁是一个高杠杆行业。根据海外第三方数据，预计从2024年到2029年，全球在AI硬件及计算中心建设上的累计资本开支将高达1.1万亿美元；到2029年，全球市场未清偿的AI专项债务预计将达到7.1万亿美元，规模仅次于美国约13万亿美元的住房抵押贷款ABS市场。国内市场同样呈现高杠杆态势。在融资模式上，国内主要有三种路径：一是基于硬件设备的融资租赁，租期通常为3到5年，算力租赁公司按月偿付租金；二是算力包销合同，这种模式在海外更为普遍，即头部的算力租赁公司与下游客户签订长期包销合同，并以此作为抵押进

行融资，实质上是利用了客户的信用；三是基于算力收益权的融资，这适用于业务模式较为成熟的算力租赁公司，根据其未来收益的测算进行抵押。总体而言，国内融资目前主要依赖设备融资租赁，而海外则更倾向于利用下游客户的长期合同作为融资凭证。

在算力租赁的融资过程中，中美两国的银行分别扮演什么角色？由此衍生出哪些不同的金融创新模式？

在算力租赁融资中，美国银行更多扮演交易组织者的角色，其风险主要在表外；而国内银行则主要是主要出资人，风险集中在银行体系内。基于角色的不同，两国形成了各具特色的金融创新模式。在美国，银行的创新方式主要有：第一，通过资产证券化或私募信贷分销，将GPU算力集群或长约收益打包成ABS产品，由私募信贷基金接盘，实现表内风险出表。第二，引入算力残值保险，联合保险机构和设备原厂锁定GPU的处置残值，以对冲资产价值波动风险。第三，通过契约驱动或衍生品对冲，引入电费、算力差价等衍生品，实现现金流的精准对冲。在国内，银行主导的创新模式包括：第一，推动智算中心公募REITs试点，将算力园区或数据中心纳入REITs的底层资产范畴。第二，推动算力收益权或Token流水质押，将算力服务合同、Token流水以及政府发放的算力券等纳入合格担保品范围。第三，通过政企协同、国资增资或地方专项债等方式进行融资。其中，若能获得地方政府专项债资金，其融资成本会显著低于市场平均水平，从而对公司的利润产生较大影响。

目前在北美和A股市场，有哪些值得关注的算力租赁相关标的？

在北美市场，较为关注的头部公司是CoreWeave和Lambda Labs。此外，由加密货币矿工转型而来的Iridian也值得注意，该公司凭借电力资源优势成功转型，在2025年实现了高速增长，并获得了微软等公司的订单，其成功也带动了行业内其他矿工向算力租赁转型。在A股市场，算力租赁公司众多，其中协创、利通、红警、高乐以及软通动力等被认为是未来在业务转型和规模扩张方面具备较大潜力的标的。

公众号：