西南证券
SOUTHWEST SECURITIES

专题报告

Deepseek引爆通信产业新机遇

西南证券研究发展中心
通信研究团队
2025年2月

核心要点

- **DeepSeek通过创新算法使推理效率大幅优化，大幅降低了应用成本。**DeepSeek-V3的训练成本仅为2.788M H800 GPU小时，同时其支持FP8混合精度训练，并针对训练框架进行了全面优化，以实现加速训练和降低GPU内存使用，通过算法、框架和硬件的共同设计，克服了跨节点MoE训练中的通信瓶颈，显著提高了训练效率并降低了训练成本。 DeepSeek每百万输入tokens成本为0.55美元，每百万输出tokens成本为2.19美元，相较于ChatGPT O1模型，输入和输出成本均降低了96%。DeepSeek-V3采用了多头潜在注意力（Multi-head Latent Attention，MLA）和DeepSeekMoE架构，显著提高了推理速度和显存利用率，能够在保持模型性能的同时实现高效的训练和推理。

- **DeepSeek从成本端和技术端对垂类AI小模型（AI Agent）带来了直接催化。**从成本端看，更低的推理成本降低了垂类AI Agent的开发成本，极大刺激了各行业的企业智能化需求。技术端看，Deepseek在自然语言理解、多模态交互等底层技术上的突破直接降低了垂直领域小模型的技术门槛，其开源的分布式训练框架等技术能够被小模型复用。同时，Deepseek的模型知识蒸馏等压缩技术使小模型既能继承大模型能力，又保持轻量化特性。对数据实时性敏感的垂类AI agent需要在感知端和云端快速传递数据，对低时延高带宽网络提出要求，同时小模型下沉到中小企业，进一步带来了网络通信基础设施需求，对交换机、边缘计算设备、5G切片等带来新需求。

- **光模块等需求来源从训练转向推理，带来多场景适配需求。**虽然单次训练任务的算力需求降低，但模型轻量化可能推动分布式训练和边缘计算的普及，导致数据中心内部短距连接需求从集中式超算集群转向更分散的节点间通信。机架内光模块对于高密度计算仍需要低延迟、高带宽的互连，800G模块需求可能受分布式架构的推动；而在边缘场景，短距光模块在边缘服务器的部署比例可能上升，但单点用量低于传统超算中心。同时，技术替代效应强于需求收缩，CPO的核心价值在于解决传统可插拔光模块的功耗和密度瓶颈，即使算力需求下降，但对于能效比要求、空间压缩要求、降低成本要求仍可能驱动其渗透率提升。

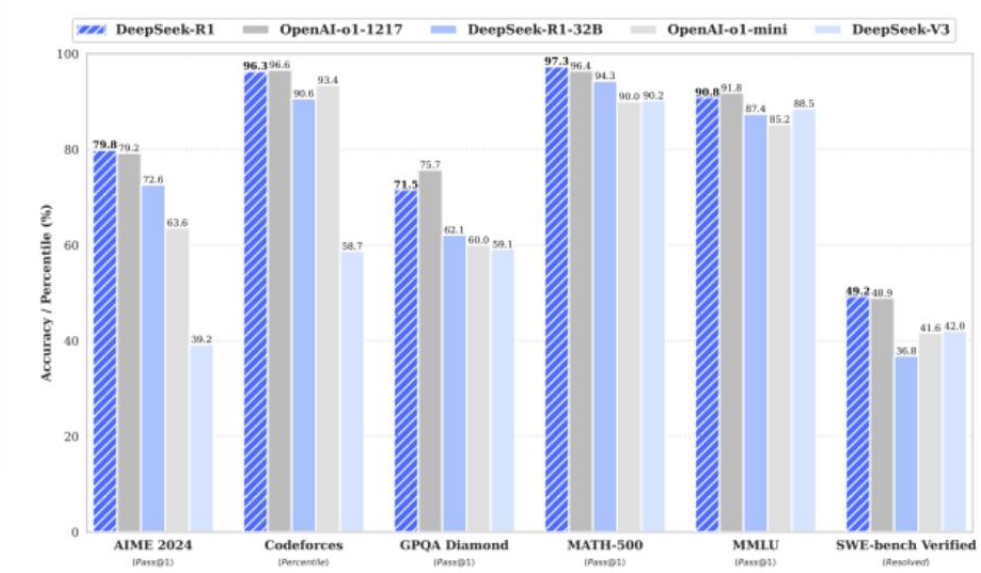
- **风险提示：**AI建设不及预期；上游资本开支不及预期等。

目 录

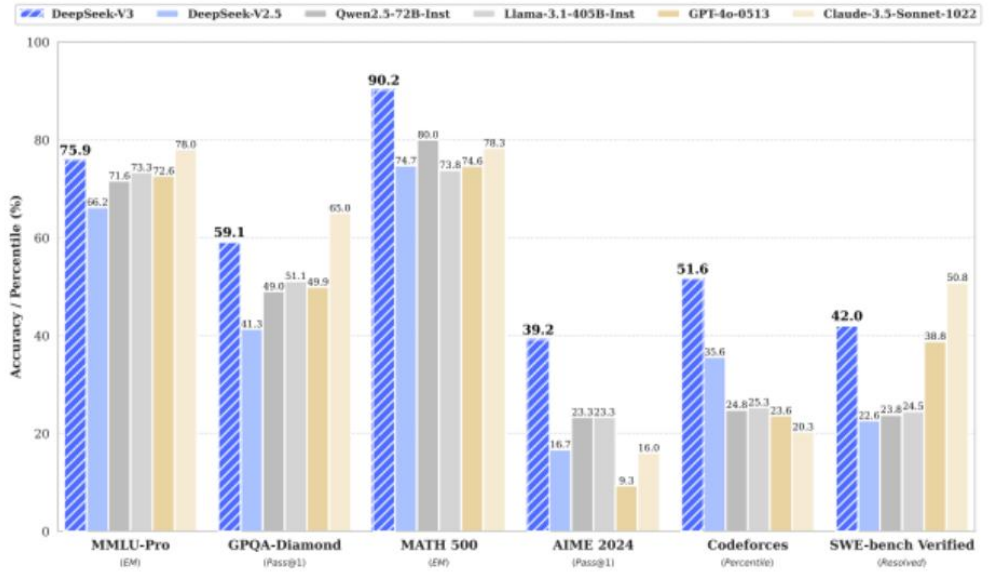
DeepSeek与开源模型的“降本增效”革命

- **DeepSeek大幅降低了应用成本。**DeepSeek-V3的训练成本仅为2.788M H800 GPU小时，同时其支持FP8混合精度训练，并针对训练框架进行了全面优化，以实现加速训练和降低GPU内存使用，通过算法、框架和硬件的共同设计，克服了跨节点MoE训练中的通信瓶颈，显著提高了训练效率并降低了训练成本。 DeepSeek每百万输入tokens成本为0.55美元，每百万输出tokens成本为2.19美元，相较于ChatGPT O1模型，输入和输出成本均降低了96%。
- **DeepSeek通过创新算法使推理效率大幅优化。**DeepSeek-V3采用了多头潜在注意力（Multi-head Latent Attention，MLA）和DeepSeekMoE架构，显著提高了推理速度和显存利用率，能够在保持模型性能的同时实现高效的训练和推理。

DeepSeek-R1与各类大模型性能比较



DeepSeek-V3与各类大模型性能比较



数据来源：DeepSeek，西南证券整理

DeepSeek与开源模型的“降本增效”革命

- **MLA架构能够大幅提升模型推理效率。**MLA（Multi-head Latent Attention）跨层注意力特征融合架构架构是DeepSeek模型中的一种注意力机制优化技术，通过低秩联合压缩注意力键（Key）和值（Value），显著降低了推理过程中的KV缓存，同时保持了与标准多头注意力（MHA）相当的性能。MLA架构在保持模型性能的同时，通过压缩技术减少了内存占用和计算量，从而提高了模型的推理效率。
- **MoE稀疏化能够控制激活参数数量，提升模型计算效率。**MoE（Mixture of Experts）通过将模型划分为多个“专家”模块，每个专家专注于处理特定的任务或数据子集。在训练和推理过程中，只有部分专家被激活，从而减少了不必要的计算。MoE架构能够显著降低计算开销，提高模型的训练和推理效率。此外，MoE架构还具有高度的可扩展性，通过增加专家的数量，可以进一步提升模型的性能，而不会显著增加计算成本。

MoE系数模型效率提升

指标	稠密模型	MoE稀疏模型	提升幅度
单样本计算量	100%参数参与	10%-30%参数参与	3-10倍
训练显存占用	全参数存储	仅激活专家存储	降低60%
千卡集群利用率	45%-50%	60%-65%	+15个百分点

MLA架构带来的效率优化

传统Transformer	MLA架构创新点	效果提升	传统Transformer
单层自注意力	跨层注意力门控	上下文理解+23%	单层自注意力
固定参数交互	动态参数共享机制	模型容量140%	固定参数交互
独立位置编码	层级相关位置偏置	长文本处理13.5x	独立位置编码

MLA架构带来的效率优化

评估维度	传统注意力架构	MLA+MoE架构	超额收益来源
资本效率	1产出需0.8算力投入	1产出需0.3算力投入	毛利率提升空间+25pp
技术代差	6-12个月迭代周期	2-4周架构微调能力	先发优势溢价+40%
生态价值	封闭式API模式	开源框架+开发者生态	网络效应乘数3x
风险敞口	依赖英伟达生态	国产化技术栈自主可控	地缘政治风险-70%

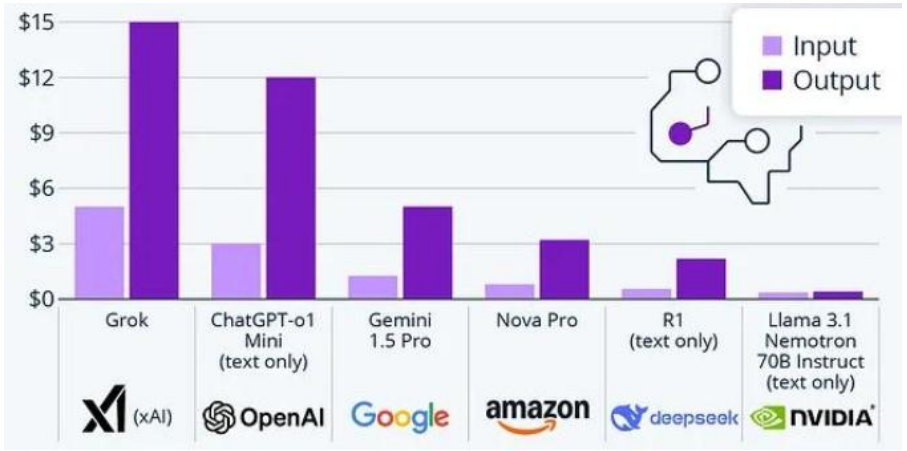
数据来源：DeepSeek，西南证券整理

目 录

垂直小模型多点开花

- **DeepSeek从成本端和技术端对垂类AI小模型（AI Agent）带来了直接催化。**从成本端看，DeepSeek推理成本大幅降低，DeepSeek每百万输入tokens成本为0.55美元，每百万输出tokens成本为2.19美元，相较于ChatGPT O1模型，输入和输出成本均降低了96%。更低的推理成本降低了垂类AI Agent的开发成本，极大刺激了各行业的企业智能化需求。技术端看，Deepseek在自然语言理解、多模态交互等底层技术上的突破直接降低了垂直领域小模型的技术门槛，其开源的分布式训练框架等技术能够被小模型复用。同时，Deepseek的模型知识蒸馏等压缩技术使小模型既能继承大模型能力，又保持轻量化特性。
- **垂类数据集结合DS大模型，形成强大数据飞轮。**Deepseek通过与各垂直领域合作获取的行业数据，反哺其基座模型优化，而基座模型优化又进一步降低了开发垂类模型的行业数据的需求和成本，形成“大模型增强-小模型易用性提升”的正向循环。

各大模型tokens成本比较



部分垂类大模型示意

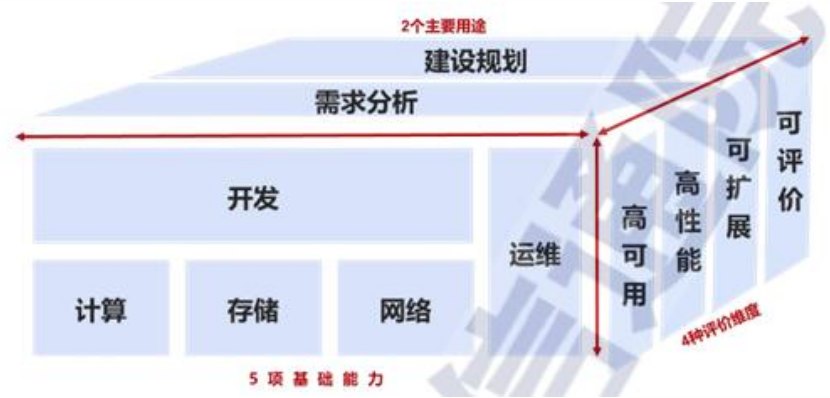
行业	垂类模型	简介
零售	微盟导购任务AI+	集成DeepSeek，帮助零售企业自动化、智能化规划导购任务，为导购的用户运营和销售转化提供方向
	京东公有云在线部署	上线 DeepSeek-R1 和 DeepSeek-V3 模型，支持公有云在线部署、专混私有化实例部署两种模式，供用户按需部署，快速调用，进行智能客服，提升电商运营效率
汽车	东风汽车DeepSeek全系列大语言模型	接入DeepSeek大模型，实现智能座舱、交互语音、自动驾驶辅助等功能，提升驾驶体验和自动化水平
金融	苏商银行DeepSeek系列技术	结合DeepSeek模型优化金融服务，如智能投顾、风险评估、客服自动化等，提升客户体验并优化运营效率
	诺安基金诺安AI助手	完成DeepSeek金融大模型的本地化部署，推出基于主流AI开源框架自主研发的“诺安AI助手”，于投研分析、客户服务、风险管控等核心业务场景启动试点应用
医疗健康	医渡云"AI医疗大脑"YiduCore	评估临床辅助诊断、医学影像分析、健康管理等领域，为研究、诊疗及公共健康三大类场景提供强有力的智能化支持，提高医疗服务标准化水平
	丁香园智能医学问答AI	集成DeepSeek，优化医生与患者的沟通效率，提供医学知识查询、病情分析等功能
高精尖制造	海光信息	完成DeepSeek V3和R1模型与海光DCU的适配，海光DCU（深度计算单元）基于高性能GPGPU架构，支持FP32/FP16高精度计算，已在金融、医疗、政务等领域实现规模化应用
教育	沪江网校AI智能助教	基于DeepSeek大模型，实现个性化学习推荐、智能答疑、作业批量修改等功能，提升在线教育体验

数据来源：Statista，微盟，东风汽车，苏商银行，诺安基金，丁香园，海光信息，沪江网校，西南证券整理

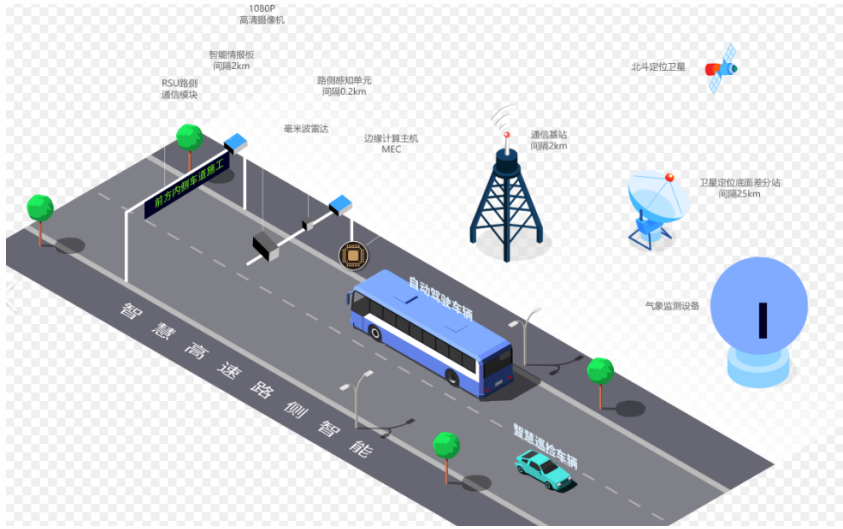
垂直小模型带来通信基础设施需求

- **对数据实时性敏感的垂类AI agent需要在感知端和云端快速传递数据，对低时延高带宽网络提出要求。**例如在自动驾驶领域，Agent需要实时获取车辆周围环境的大量数据，如路况、交通信号、其他车辆和行人的位置等，以便做出快速准确的决策。这要求5G网络的端到端时延必须小于20ms，以确保数据的实时传输和处理。
- **大量的数据处理也带来了算力分布的重构。**医疗影像领域需要处理大量的高分辨率图像数据，传统的集中式云计算模式在处理这些数据时存在延迟和带宽瓶颈，因此，边缘计算节点可以将部分计算任务从云端转移到靠近数据源的地方，从而减少数据传输的延迟，提高数据处理的效率。
- **工业数据也对网络可靠性提出要求。**工业质检领域对通信基础设施的网络可靠性提出了极高的要求，质检模型需要实时传输和处理大量的高清图像和视频数据，任何网络中断或数据丢失都可能导致质检过程的中断，一般工业确定性网络的需要可用性达到99.999%，以实现数据的确定性传输。

网络是大模型的重要基础能力



车联网边缘计算示意图

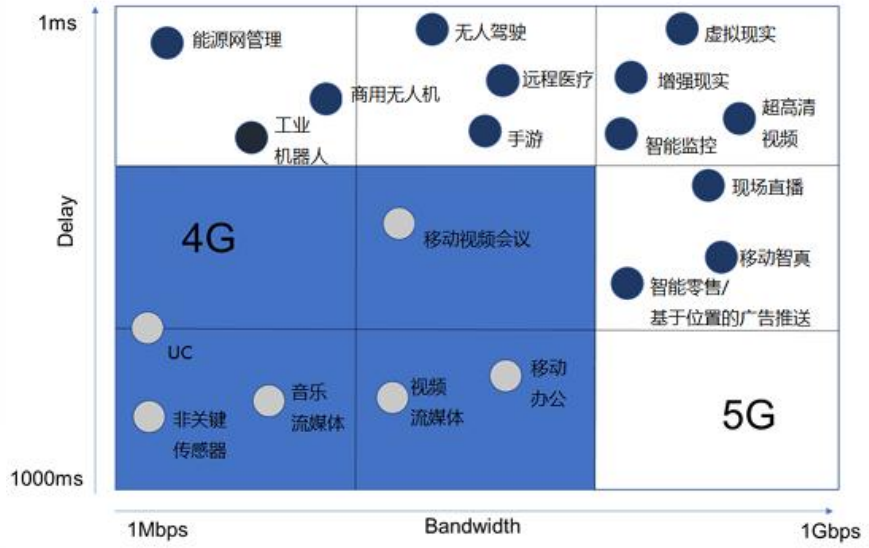


数据来源：中国信通院，西南证券整理

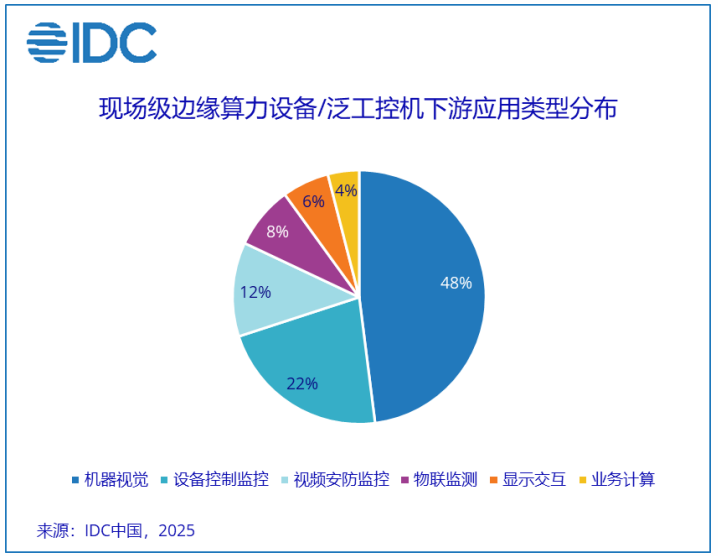
网络架构革命驱动基础设施升级

- **5G增强型基站爆发：**边缘计算将数据处理和分析的能力从云端向网络边缘转移，靠近数据源或用户端，5G增强型基站作为5G网络的重要组成部分，其高速率、低时延、大连接的特性为边缘计算提供了强大的网络支持，使得边缘计算能够更好地发挥其优势。在边缘计算场景下，边缘AI要求网络切片能力提升300%，AAU设备需支持动态算力分配，华为、中兴均已重点布局。
- **边缘DC规模化部署：**据IDC发布的《中国设备现场级边缘算力设备 / 泛工控机市场份额，2023》，随着边缘计算的兴起和发展，全球边缘计算支出预计将在2028年达到3780亿美元，48%的边缘计算设备应用于机器视觉领域，22%和12%的设备应用于设备控制监控和视频安防监控领域。同时，边缘计算算力需求也对设备的功耗和冷却提出了新要求，24年单台边缘服务器功率密度突破30kW/机架，预计液冷渗透率在未来三年内将从5%升至35%。

各应用对于速率和时延需求分布



边缘计算设备应用领域分布

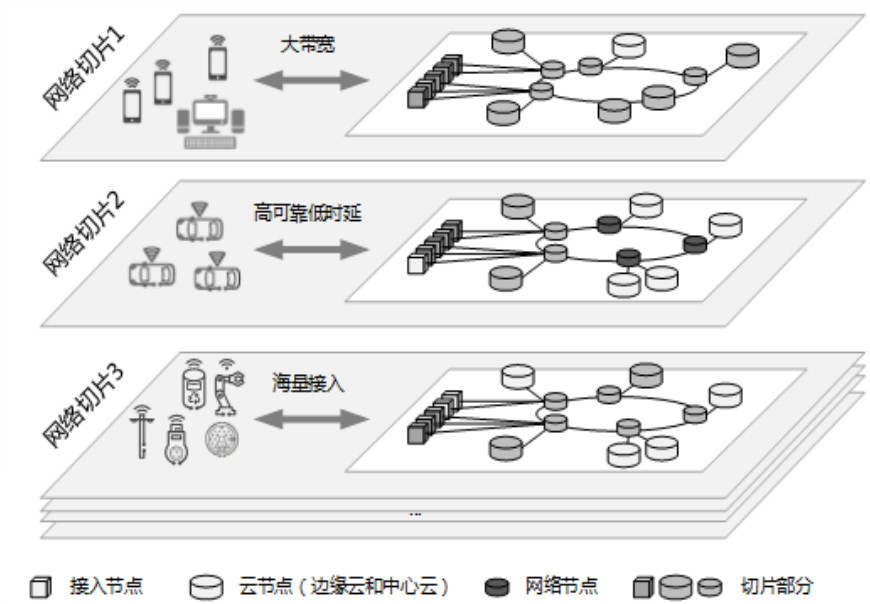


数据来源：OVUM，IDC，西南证券整理

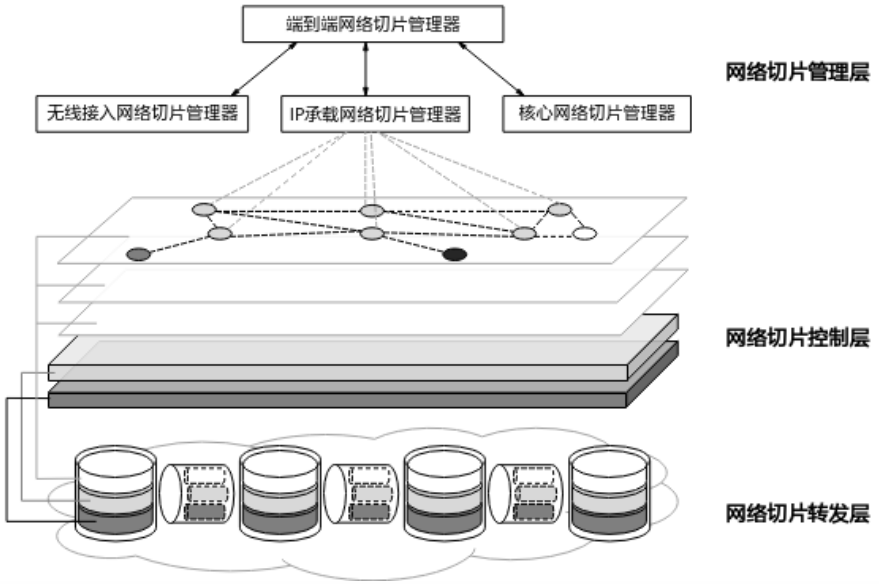
运营模式向算力服务转型

- **网络切片即服务（NSaaS）**：网络切片是5G网络的一项关键技术，它将物理网络切割成多个虚拟网络，每个切片具备独立的逻辑功能和性能特征，可按需定制，满足不同业务需求。边缘计算与网络切片结合，可实现网络资源的灵活分配和管理。AI Agent可以按需调用切片资源，根据不同应用对网络带宽、时延的不同要求，利用网络切片技术，可为各应用创建专属切片，结合边缘计算的数据处理能力，实现生产过程的高效协同和优化。
- **算力交易平台**：算力交易平台是连接算力供需双方的中介，提供算力资源的交易、调度和管理服务。它将分散的算力资源整合，形成算力资源池，根据用户需求进行灵活分配和调度。边缘计算的算力资源分布广泛且异构，算力交易平台可有效整合这些资源，提高资源利用率。

网络切片示意图



网络切片架构示意图



数据来源：华为，西南证券整理

垂直行业应用价值裂变

DeepSeek的效率升级对各类垂直行业的ai应用均带来了价值倍增效应

- **工业质检领域：**通过在生产现场部署边缘计算设备，可以实时处理和分析来自生产线的大量数据，如高清图像、传感器数据等，实现对产品质量的快速检测和缺陷识别。如映翰通在 EC5000系列边缘计算机上完成 DeepSeek R1 蒸馏模型的本地部署。这一成果验证了轻量级边缘设备（如 EC5000）在 AI 推理任务中的强大潜力。相较于传统云端部署，边缘端计算无需依赖高算力服务器，可在低功耗环境下实现实时推理，为工业质检、智慧交通、远程医疗等领域提供更灵活、安全、高效的AI 解决方案。
- **车联网领域：**通过在道路侧部署边缘计算设备，可以实时处理和分析来自车辆的大量数据，如车辆位置、速度、行驶状态等，实现对交通流量的优化管理和车辆的智能控制。
- **政务领域：**边缘计算在政务领域的应用主要体现在数据安全和隐私保护方面。通过在本地部署边缘计算设备，可以确保敏感数据不外泄。
- **金融领域：**通过在本地部署边缘计算设备，可以实时处理和分析金融交易数据，实现对风险的快速识别和控制，同时提供更加个性化的金融服务。目前已有20家券商已官宣完成DeepSeek的接入或本地化部署，覆盖合规问答、业务办理指引、知识查询、投研分析等多个场景。
- **能源领域：**边缘计算在能源领域的应用主要体现在能源管理和设备监控方面。通过在能源生产现场部署边缘计算设备，可以实时监控能源设备的运行状态，实现对能源的高效管理和优化调度，提高能源利用效率。

映翰通边缘网关计算机

数据来源：映翰通，西南证券整理

目 录

光通信：从“速度崇拜”到“场景适配”

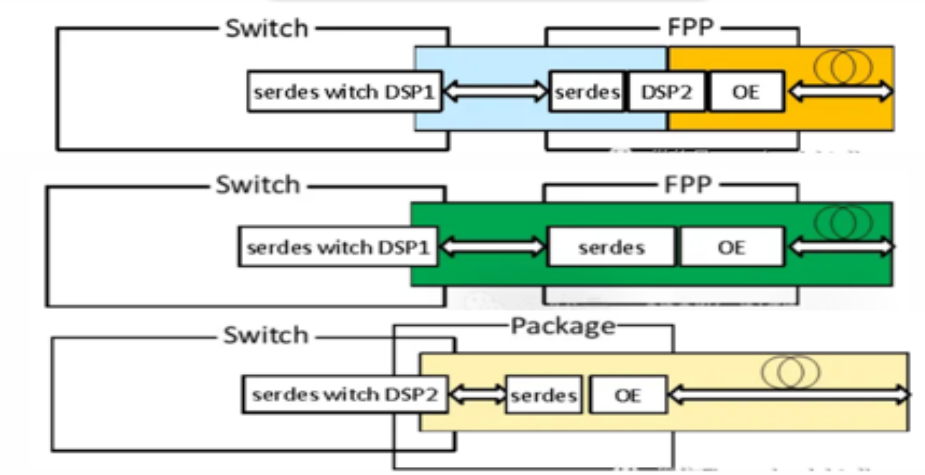
虽然单次训练任务的算力需求降低，但模型轻量化可能推动分布式训练和边缘计算的普及，**导致数据中心内部短距连接需求从集中式超算集群转向更分散的节点间通信。**

- 机架内光模块：高密度计算仍需要低延迟、高带宽的互连，800G模块需求可能受分布式架构的推动。
- 边缘场景：短距光模块在边缘服务器的部署比例可能上升，但单点用量低于传统超算中心。

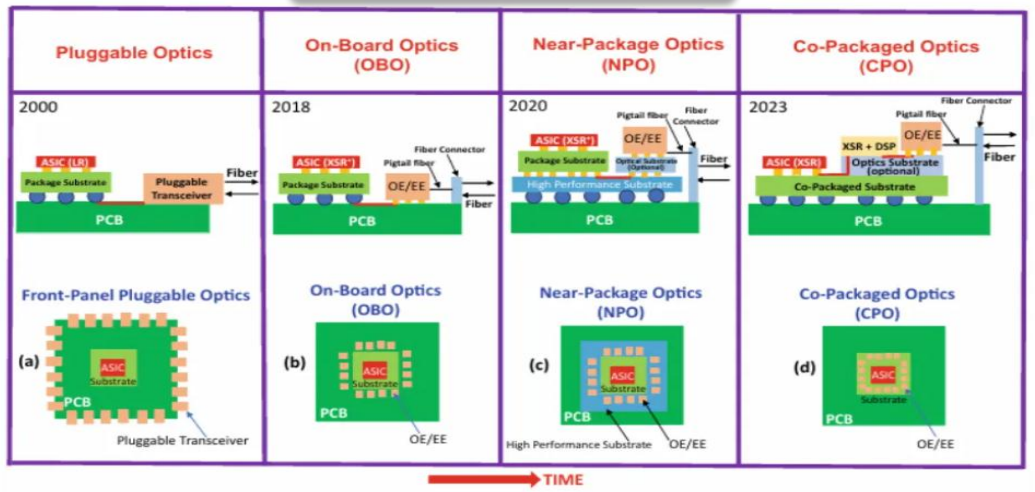
技术替代效应强于需求收缩。CPO的核心价值在于解决传统可插拔光模块的功耗和密度瓶颈。即使算力需求下降，但以下因素仍可能驱动其渗透率提升：能效比要求：AI芯片的功耗优化倒逼互连技术提升能效，CPO的功耗优势（降低30-50%）可能成为刚需。

- 空间压缩：轻量化模型可能催生紧凑型训练节点（如多芯片封装），CPO的集成优势更适配高密度封装。
- 降低成本：硅光方案成本优势凸显（较EML方案低40%），Lumentum、Intel硅光产线利用率提升，传统厂商（光迅科技）加速转型。

FPP/LPO/CPO示意图



FPP/NPO/CPO示意图

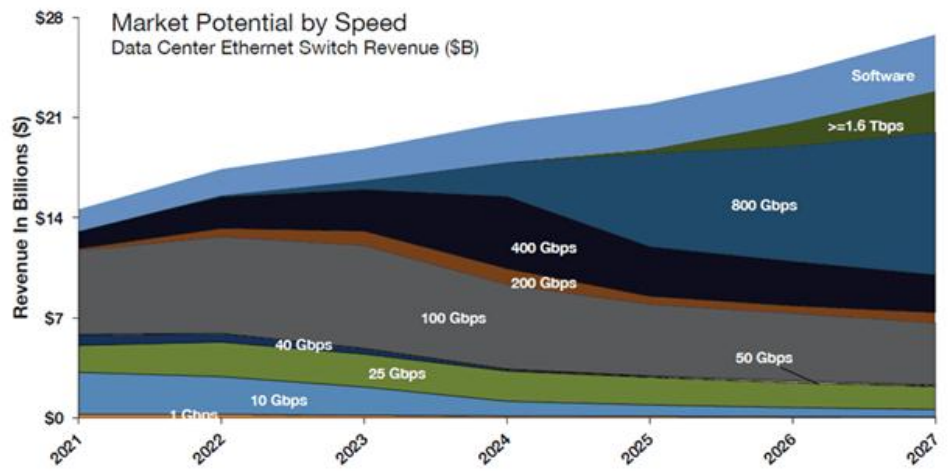


数据来源：CSDN，西南证券整理

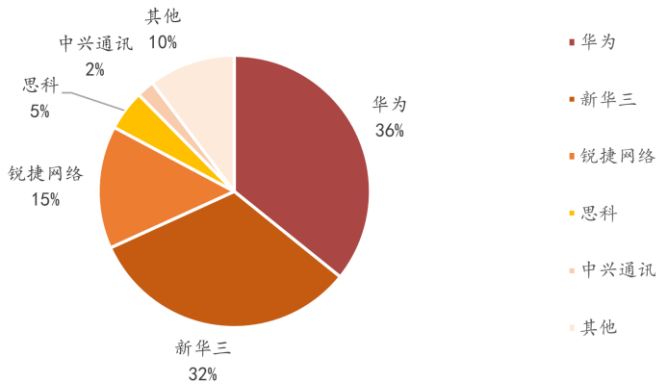
算力及网络设备：从集中到分布

- **边缘与中小企业交换机市场有望爆发。**随着轻量化AI模型的推广， AI推理能力正逐渐扩展到工厂、医院、零售等边缘场景，中小企业得益于低成本算力的普及，现在有能力自行部署私有化的AI集群，这不仅降低了对外部云服务的依赖，还增强了数据的安全性和隐私保护。随着边缘推理场景的扩张和中小企业私有化部署的普及，将会带来园区交换机和数据中心交换机需求提升，而中小企业对于性价比和能耗关注度高，国内交换机企业有望受益。
- **不同垂类行业的可定制化需求加速软硬件解耦。**软硬件解耦使得交换机的硬件和软件可以独立发展和升级，硬件部分可以采用标准化的组件，降低成本和提高兼容性；软件部分可以根据行业需求进行定制化开发，提供更灵活和高效的功能。解耦推进对于国内交换机和上游组件厂商有较好的催化，相关细分产业的公司可以通过在细分领域的竞争力获取订单和份额。

高速交换机需求提升



国内交换机市场份额（2024）

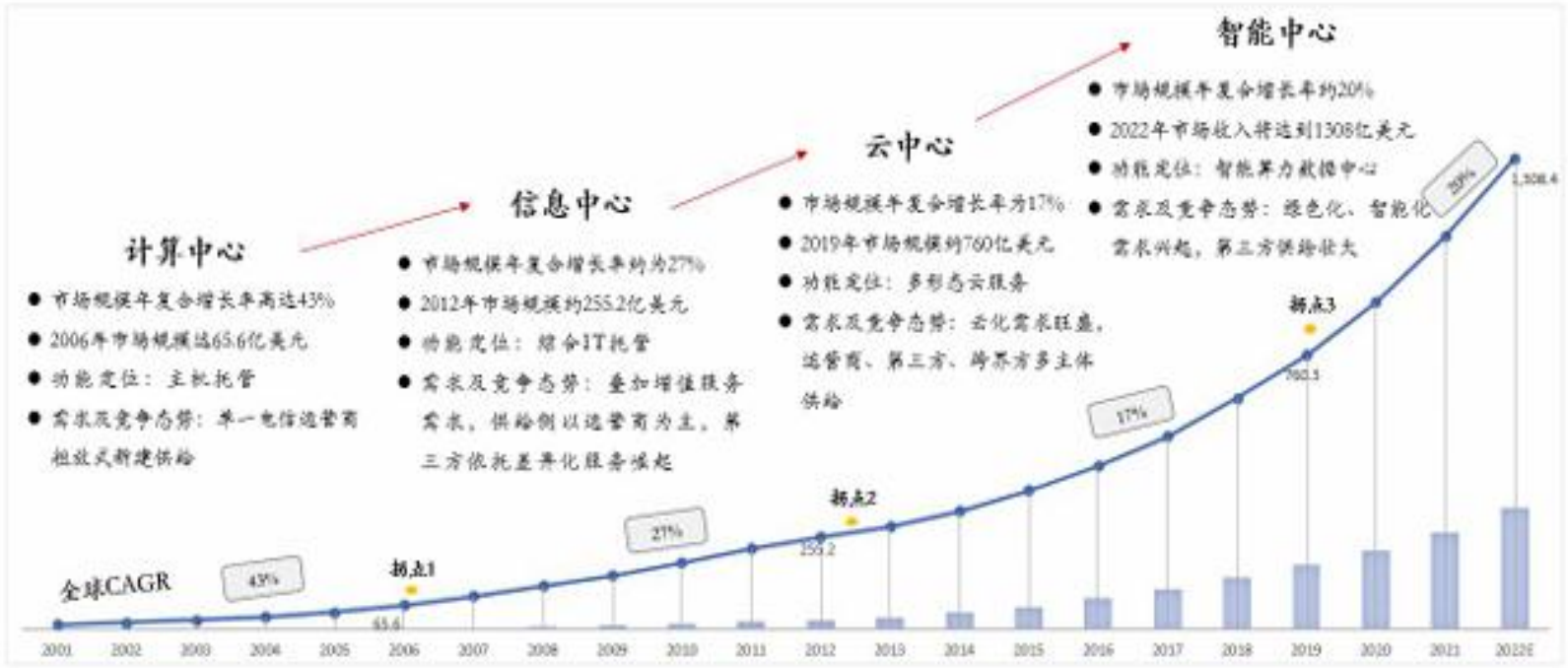


数据来源：IDC，中商产业研究院，西南证券整理

智算中心：AI应用下沉，边缘部署需求增强

- 随着AI应用下沉，企业还关注将推理部署到边缘以降低时延和带宽占用。运行大型模型的精简版本于本地设备已成为趋势，使服务更实时可靠，并降低云端压力。例如，DeepSeek-R1等新一代开源模型宣称实现小型化部署，能在笔记本乃至嵌入式设备运行强大的推理功能。这意味着智算中心不仅需提供云端算力，还可能扮演边缘AI的训练支撑与协同角色，为边缘设备提供预训练模型和更新支持。**在未来，中心-边缘协同的计算架构将更普遍：中心负责训练大模型和复杂推理，边缘负责本地实时推理，两者共同满足企业的AI需求。**

全球数据中心向智算中心发展



数据来源：信科院《中国数据中心产业发展白皮书（2023年）》，西南证券整理

风险提示

- AI建设不及预期;
- 上游资本开支不及预期等。

西南证券
SOUTHWEST SECURITIES

分析师：叶泽佑
执业证号：S1250522090003
电话：13524424436
邮箱：yezy@swsc.com.cn

分析师：曾庆亮
执业证号：S1250524080001
邮箱：zqlyf@swsc.com.cn

西南证券投资评级说明

报告中投资建议所涉及的评级分为公司评级和行业评级（另有说明的除外）。评级标准为报告发布日后6个月内的相对市场表现，即：以报告发布日后6个月内公司股价（或行业指数）相对同期相关证券市场代表性指数的涨跌幅作为基准。其中：A股市场以沪深300指数为基准，新三板市场以三板成指（针对协议转让标的）或三板做市指数（针对做市转让标的）为基准；香港市场以恒生指数为基准；美国市场以纳斯达克综合指数或标普500指数为基准。

公司评级	买入：未来6个月内，个股相对同期相关证券市场代表性指数涨幅在20%以上
	持有：未来6个月内，个股相对同期相关证券市场代表性指数涨幅介于10%与20%之间
	中性：未来6个月内，个股相对同期相关证券市场代表性指数涨幅介于-10%与10%之间
	回避：未来6个月内，个股相对同期相关证券市场代表性指数涨幅介于-20%与-10%之间
	卖出：未来6个月内，个股相对同期相关证券市场代表性指数涨幅在-20%以下
行业评级	强于大市：未来6个月内，行业整体回报高于同期相关证券市场代表性指数5%以上
	跟随大市：未来6个月内，行业整体回报介于同期相关证券市场代表性指数-5%与5%之间
	弱于大市：未来6个月内，行业整体回报低于同期相关证券市场代表性指数-5%以下

分析师承诺

报告署名分析师具有中国证券业协会授予的证券投资咨询执业资格并注册为证券分析师，报告所采用的数据均来自合法合规渠道，分析逻辑基于分析师的职业理解，通过合理判断得出结论，独立、客观地出具本报告。分析师承诺不曾因，不因，也将不会因本报告中的具体推荐意见或观点而直接或间接获取任何形式的补偿。

重要声明

西南证券 SOUTHWEST SECURITIES

西南证券研究发展中心

西南证券研究发展中心

上海

地址：上海市浦东新区陆家嘴21世纪大厦10楼

邮编：200120

北京

地址：北京市西城区金融大街35号国际企业大厦A座8楼

邮编：100033

深圳

地址：深圳市福田区益田路6001号太平金融大厦22楼

邮编：518038

重庆

地址：重庆市江北区金沙门路32号西南证券总部大楼21楼

邮编：400025

西南证券机构销售团队

区域	姓名	职务	手机	邮箱	姓名	职务	手机	邮箱
上海	蒋诗烽	总经理助理/销售总监	18621310081	jsf@swsc.com.cn	欧若诗	销售经理	18223769969	ors@swsc.com.cn
	崔露文	销售副总监	15642960315	clw@swsc.com.cn	李嘉隆	销售经理	15800507223	ljlong@swsc.com.cn
	李煜	高级销售经理	18801732511	yfliyu@swsc.com.cn	龚怡芸	销售经理	13524211935	gongyy@swsc.com.cn
	田婧雯	高级销售经理	18817337408	tjw@swsc.com.cn	孙启迪	销售经理	19946297109	sqdi@swsc.com.cn
	张玉梅	销售经理	18957157330	zymyf@swsc.com.cn	蒋宇洁	销售经理	15905851569	jyj@swsc.com.c
	魏晓阳	销售经理	15026480118	wxyang@swsc.com.cn				
北京	李杨	销售总监	18601139362	yfly@swsc.com.cn	张鑫	高级销售经理	15981953220	zhxin@swsc.com.cn
	张岚	销售副总监	18601241803	zhanglan@swsc.com.cn	王一菲	高级销售经理	18040060359	wyf@swsc.com.cn
	杨薇	资深销售经理	15652285702	yangwei@swsc.com.cn	王宇飞	高级销售经理	18500981866	wangyuf@swsc.com
	姚航	高级销售经理	15652026677	yhang@swsc.com.cn	马冰竹	销售经理	13126590325	mbz@swsc.com.cn
广深	郑龑	广深销售负责人	18825189744	zhengyan@swsc.com.cn	杨举	销售经理	13668255142	yangju@swsc.com.cn
	杨新意	广深销售联席负责人	17628609919	yxy@swsc.com.cn	陈韵然	销售经理	18208801355	cyryf@swsc.com.cn
	龚之涵	高级销售经理	15808001926	gongzh@swsc.com.cn	林哲睿	销售经理	15602268757	lzr@swsc.com.cn