北美大厂AI收费模式变革：从Tokens到PTU的演进与投资机遇

行业研究·行业专题

计算机·人工智能

投资评级：优于大市

证券分析师：熊莉

021-61761067

xiongli1@guosen.com.cn

S0980519030002

摘要

国信证券 GUOSEN SECURITIES

- **事件：海外云厂数据中心AI收费方式从Tokens到PTU的变化，是行业为适应AI应用增长和成本管理需求而进行的调整。**随着AI Agent等应用兴起，Token消耗量呈指数级增长，超大型客户的Token账单失控，促使云厂商推动计费模式转型。PTU模式成为超大型客户的主流选择，行业整体呈现“ PTU + Token”的格局，并逐步向更高比例PTU演进。
- **影响一：模式切换初期ROI阶段性触底。**定价复杂性带来短期利润波动，但这是行业成熟的必经之路，市场预期已充分消化该利空。
- **影响二：为客户提供更优服务，提升客户粘性。**PTU/MaaS模式提供可预测成本，深度绑定大客户，长期将显著提升AI用量与粘性，构建更可持续的收入增长引擎。
- **影响三：毛利率从“波动”到“稳健”变化。**短期来看，该收费策略的转变带来云业务毛利率旅游下；但随着资源利用率的提升及合同稳定性的增强，远期来看，整体毛利率结构从高弹性向高韧性转变，盈利质量更佳。
- **影响四：定价变革或可带来新的投资机遇。**云厂定价模式目前正处于变革拐点期，预计未来一年内定价变革影响消化完毕，订单长协化趋势明显，收入与利润增长确定性显著提升，或可带来新的投资机遇。
- **风险提示**：AI应用落地不及预期、市场需求不及预期、行业竞争加剧、宏观经济波动。

目录

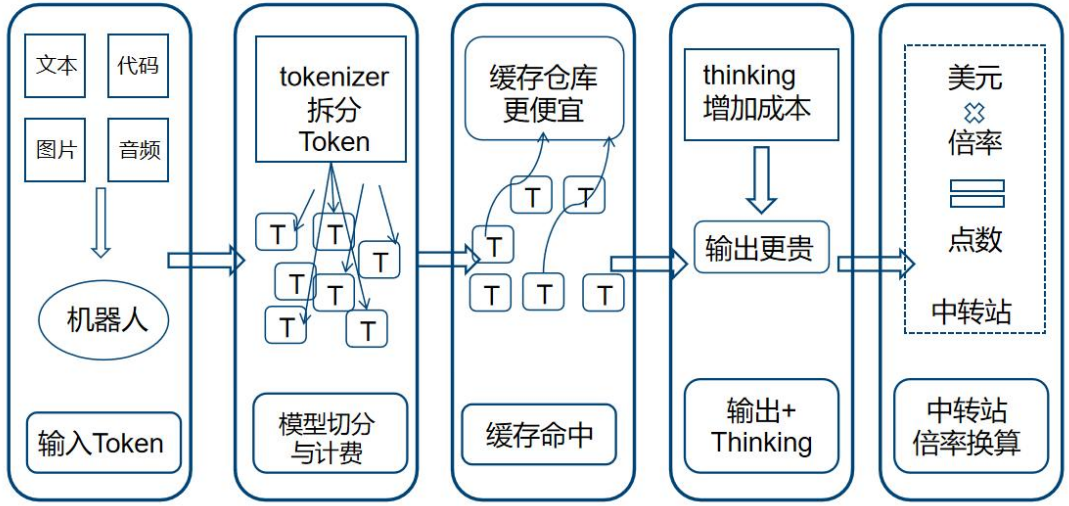
Tokens计费模式不利于企业长期规划预算

国信证券 GUOSEN SECURITIES

➢ **Tokens计费模式存在局限性，在AI快速发展的背景下，成为企业规模化应用的障碍。**

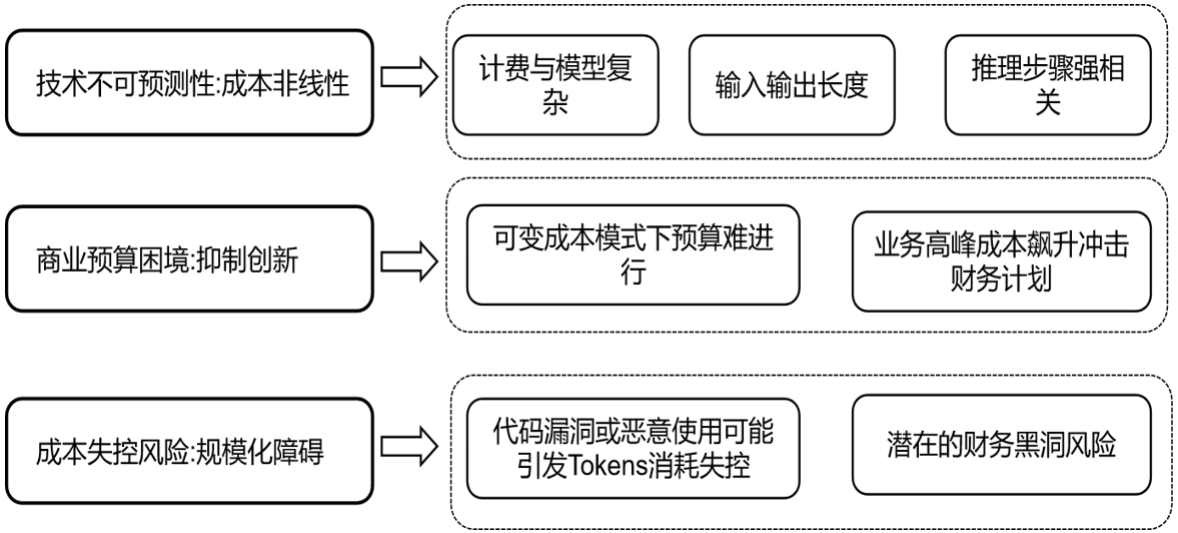
- Token计费模式：以实际消耗的Token数量为计价单位，用户按调用AI模型时产生的输入/输出Token量付费。
- **特点：**适用于中小客户或按需使用场景，灵活性高，但成本随使用量波动，超大型客户若Token消耗量巨大，账单可能失控。
- **局限：**难以预测长期成本，不利于企业规划预算，且在高频调用或复杂任务场景下，成本可能迅速攀升。**技术不可预测性与商业预算困境共同导致客户抵制，阻碍AI业务落地。**

图1：token计费模式



资料来源：Gartner，国信经济研究所整理

图2：token计费模式的局限性



资料来源：Gartner，国信经济研究所整理

AI定价模式演进：从“按次零售”到“资源订阅”

国信证券 GUOSEN SECURITIES

- **AI定价模式演进：从“按次零售”到“资源订阅”的必然阶段。**随着AI Agent等应用兴起，Token消耗量呈指数级增长，超大型客户（如A3级别以上）的Token账单失控，促使云厂商推动计费模式转型。PTU模式成为超大型客户的主流选择。

- PTU（预配吞吐量单位）计费模式：用户预先购买一定数量的PTU，云厂商为其分配相应的计算资源（包括硬件、软件、模型等），按年/月/季固定收费，与实际Token消耗量无关。

- 特点：1）成本可预测。用户可提前锁定成本，便于企业财务规划。2）资源保障：确保用户独享计算资源，避免因共享资源导致的延迟或性能波动，适合对延迟敏感、需稳定吞吐量的场景。3）客户粘性高：通常要求用户签订长期合同（如3年），增强客户与云厂商的合作稳定性。

- 局限：若用户实际使用量低于购买的PTU，可能存在资源闲置成本；若使用量超出PTU，需额外付费或升级套餐。

图3：token计费模式向PTU计费模式的演变

资料来源：国信经济研究所整理

定价模式类比移动互联网：从0.01元/KB到“无限流量”

国信证券 GUOSEN SECURITIES

- **AI收费模式从Tokens逐步切换到PTU，这种“按次零售”到“资源订阅”，是行业经历的必然阶段。定价模式的转型类似于中国的移动互联网收费，经历了从最早的0.01元/KB到今天的包月“无限流量”，用户从“购买流量”转向”购买服务“的转变。**

- 2G时代：1994年，中国首个GSM电话拨通，正式进入2G时代。2000年以后，中国移动推出”神州行“等品牌，流量收费模式为3元/15M、5元/30M，超出部分按0.01元/KB计费。2003年中国移动在部分省市推出WAP包月套餐，每月流量不限，收费在10-20元之间，短信和通话仍然为主要业务。
- 3G时代：2009年，工信部正式发放3G牌照，我国进入3G时代，流量计费从KB转向MB。以2009年中国移动推出的无线上网卡为例：150元/月包3GB国内流量，超出后按0.1元/MB计费。随着2010年iphone 4引入国内，流量需求激增，流量单价大幅下降。
- 4G时代：2013年4G牌照发放，中国进入4G时代。2015年国家明确提出”提速降费“，流量单价大幅降低，进入包月无线流量时代。2016年11月腾讯推出”大王卡“，开启”定向免流“时代，19元包月，腾讯系APP免流，用户从”购买流量“转向”购买服务“。此后不限量套餐全面爆发。
- 5G时代：2019年6月，5G牌照正式发放，中国进入5G时代。进入”按速率分级“时代，推出如199元套餐限速1GB，299元套餐限速2GB的速率套餐，流量多少不再成为定价的核心。

图4：中国移动TD-SCDMA数据流量套餐结构

中国移动TD-SCDMA数据套餐流量结构					
套餐类型	套餐月费	包含流量(T网和切换到G网)	超出套餐部分的流量资费		每月费用封顶
			T网	切换至G网	
标准资费	0	0	0.01元/KB	0.01元/KB	
5元套餐	5元	30MB			
20元套餐	20元	50MB			1000元
50元套餐	50元	500MB	0.01元/KB		
100元套餐	100元	3GB			
200元套餐	200元	5GB			

资料来源：中国移动，国信经济研究所整理

图5：中国电信5G-A套餐资费标准

月费 (元/月)	国内流量	国内通话 (分钟)	最高下行速度	最高上行速度	其他权益
199	120GB	1000	1Gbps	200Mbps	彩话提醒、智能应答、视频彩铃等
299	180GB	1500	2Gbps	300Mbps	在199元档基础上更丰富
399	240GB	2000	3Gbps	400Mbps	权益进一步提升

资料来源：中国电信，国信经济研究所整理

目录

AI定价模式改变云厂商Capex

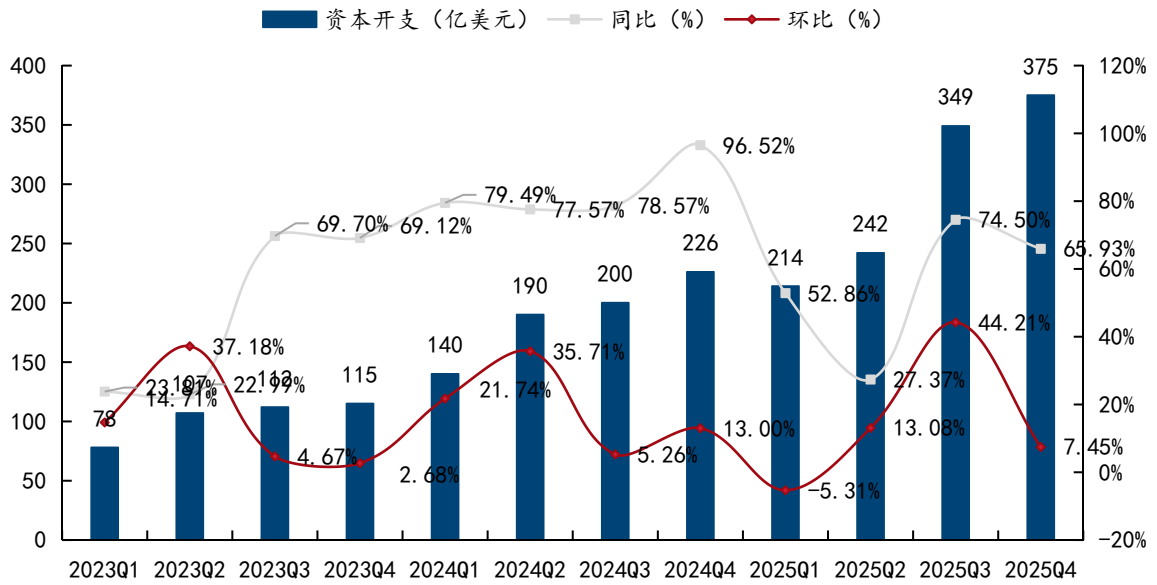
国信证券 GUOSEN SECURITIES

➢ **AI定价模式演进，会对云厂商的财务状况产生两大核心影响。在资本支出方面：**

- **Tokens模式下：厂商基于不确定需求投资，导致CAPEX呈脉冲式波动，风险很高。**在按需计费模式下，云厂商收入波动大，难以精确预测未来算力需求。这导致其在进行大规模基础设施投资时面临巨大风险：投资不足可能导致服务中断和客户流失；投资过度则会造成资源闲置和成本浪费。这种“不确定性”的投入模式财务风险极高。

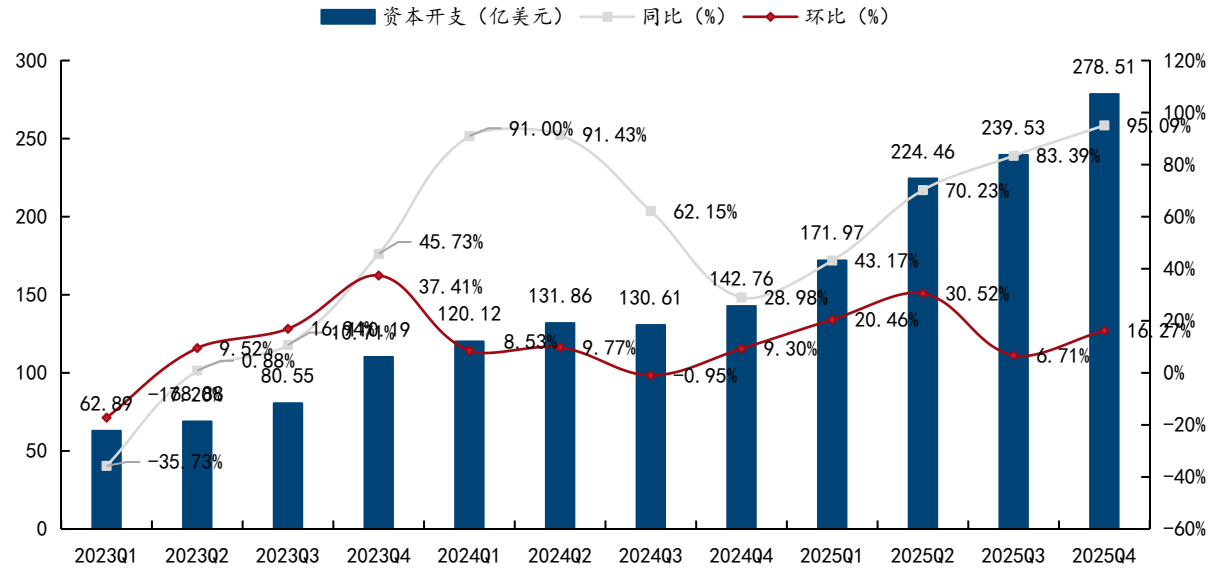
- **PTU长协模式下：CAPEX可以根据合同按需规划，变得更加平滑和可预测。**长期承诺合同提供可预测现金流。这使得厂商能：1) 基于锁定需求精确规划采购；2) 利用长单优势优化供应链；3) 锁定能源供应确保稳定。这降低了财务风险，提升了资本效率。

图6：微软资本开支同环比均提升（季度均为自然年季度，25Q3=微软FY26Q1，单位：亿美元）



资料来源：微软财报，国信经济研究所整理

图7：谷歌资本开支，同比持续提升



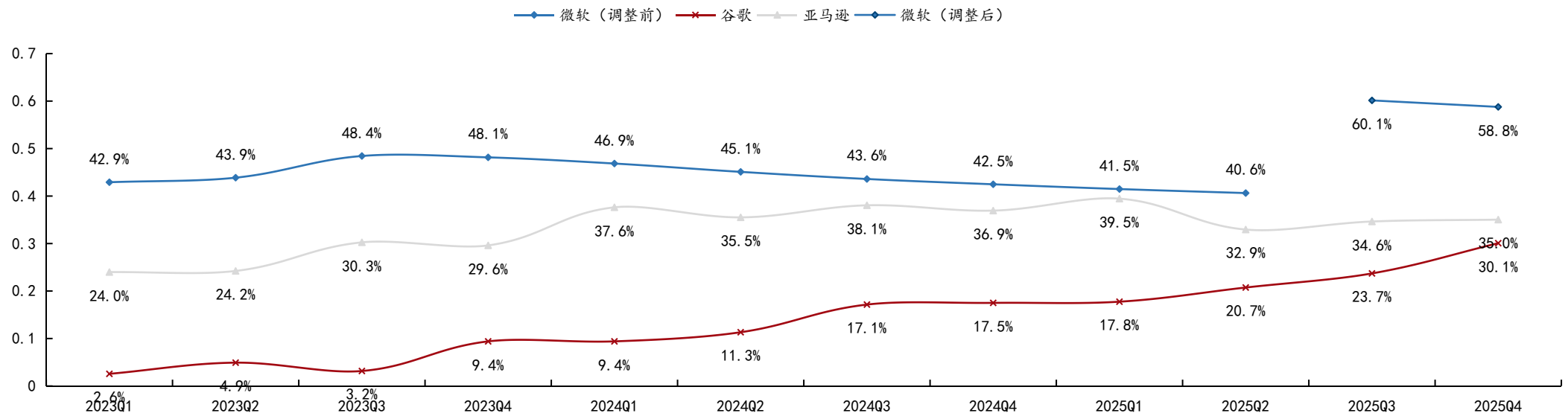
资料来源：谷歌财报，国信经济研究所整理

毛利率短期承压下降，长期实现稳定增长

- **AI定价模式演进，会对云厂商的财务状况产生两大核心影响。在毛利率方面：**Tokens模式下：毛利率差异较大，极不稳定。PTU模式下：虽然单客户利润可能不那么高，但由于资源利用率提升和客户被锁定，整体毛利率会进入一个更稳定、更健康的增长通道。

- **毛利率短期承压下降,长期客户回升+成本优化，实现稳定增长。短期：**为了吸引客户从按需付费转向长期承诺（PTU），云厂商必须提供有吸引力的折扣。因此每单位算力的毛利率会暂时下降，市场可能会因此担忧云厂商的盈利能力。**长期：**长期合同极大地增强了客户粘性，降低了流失率。同时，由于成本可预测，客户会更愿意扩大 AI 应用的规模，总用量的提升将弥补单价下降的损失。此外，可预测的需求使云厂商能优化供应链（如与芯片商谈更好价格）和运营成本（如提升数据中心利用率），可以摊薄单位成本。**定价模式转变，使得毛利率将从短期的低点回升，并进入一个比 Tokens 模式下更稳定、更健康的增长通道。**

图8：各厂商云业务毛利率或可短期承压

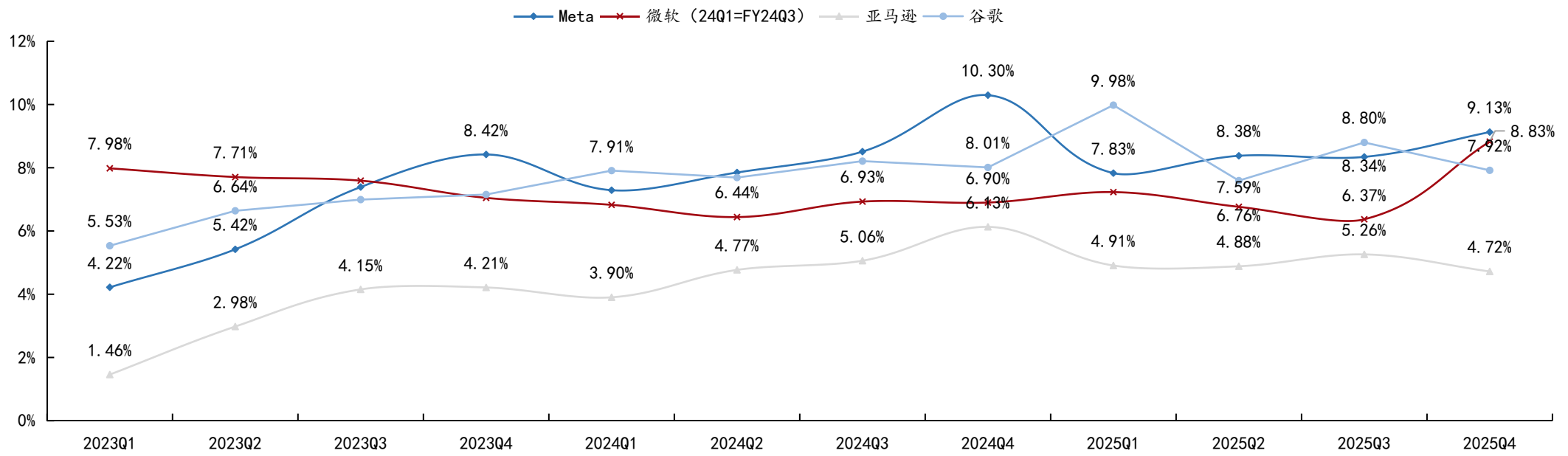


资料来源：亚马逊财报，微软财报，谷歌财报，Meta财报，国信证券经济研究所整理（注：微软25Q3之后单独披露云业务毛利率，此前仅披露利润率。计算方式为：部门营业利润/部门营收，Google和亚马逊计算方式同微软）

AI定价模式改变云厂商ROIC/毛利率

- **投入资本回报率(ROIC)方面：从历史表现来看，**受整体收入、利润增长推动，2025Q4大厂R0IC整体同比上升。**回顾近几年数据，AI需求推动各厂商资本开支大幅上行，公司的ROIC表现受到明显影响，随着AI推动收入、利润提升，整体呈现现上行趋势。**2025Q4，受AI、云服务等需求驱动.Meta、微软净利润维持同比增长，ROIC分别同比增长0.78、2.46个pct。其中Meta、微软净利润环比增速较快影响整体ROIC环比表现。
- **ROIC未来会触底反弹，或可显著提升。**长期合同让 CAPEX 从 “脉冲式” 的盲目投资转变为 “平滑式” 的精准投资，资本使用效率大大提高。此外资产周转率也有提升，同样的硬件资产能产生更多的收入，资产周转率提升。长期毛利率的稳定增长将提升净利润，也会带来ROIC的进一步提升。

图9：各厂商过去ROIC情况，整体呈现上行趋势



资料来源：亚马逊财报，微软财报，谷歌财报，Meta财报，国信证券经济研究所整理

目录

云厂商基于核心优势的差异化战略布局

国信证券 GUOSEN SECURITIES

➢ **面对 AI 定价方式从 Tokens 向 PTU / 长期承诺的转变，全球云厂商基于各自的核心优势，采取了差异化的应对策略。**

- **微软：**具备强大的企业客户基础、与 OpenAI 的深度绑定、以及 Windows、Office 365、GitHub 等核心产品构成的庞大生态系统。它不只是提供 AI 算力，而是将 AI 能力（特别是基于 GPT 的 Copilot）深度融入其核心生产力工具中。通过推出 “Azure AI 承诺计划”，鼓励企业客户签订 1-3 年的消费承诺合同，将 AI 支出锁定在其整个云生态系统内。其目标是让客户在使用 Office、Windows 等日常工具时，自然而然地消费 AI 服务，从而实现深度锁定。
- **亚马逊：**全球云市场份额第一，拥有最广泛的客户基础和最强大的基础设施，同时通过自研 Trainium 和 Inferentia 芯片获得了显著的成本优势。大力推广 “AI/ML 节省计划”（Savings Plans），为客户提供比按需付费显著优惠的价格，特别是针对使用其自研芯片的任务。AWS 的目标是利用其规模效应和成本控制能力，以最具竞争力的价格吸引客户签订长期承诺，从而巩固其市场领导地位。
- **谷歌：**在 AI/ML 领域拥有深厚的技术积累，自研的 TPU在性能上具有领先优势，同时拥有 PaLM、Gemini 等强大的自研大模型。谷歌强调其在 AI 技术上的领先地位，通过提供高性能的 TPU Pods 集群和先进的 Vertex AI 平台，吸引对性能和技术有极致要求的客户。谷歌将其 “承诺使用折扣”（CUDs）扩展至 AI 平台，旨在通过技术优势而非单纯的价格战来赢得长期合同。

图10：微软/谷歌/亚马逊等云厂的差异化布局

资料来源：微软官网，谷歌官网，亚马逊官网，国信经济研究所整理

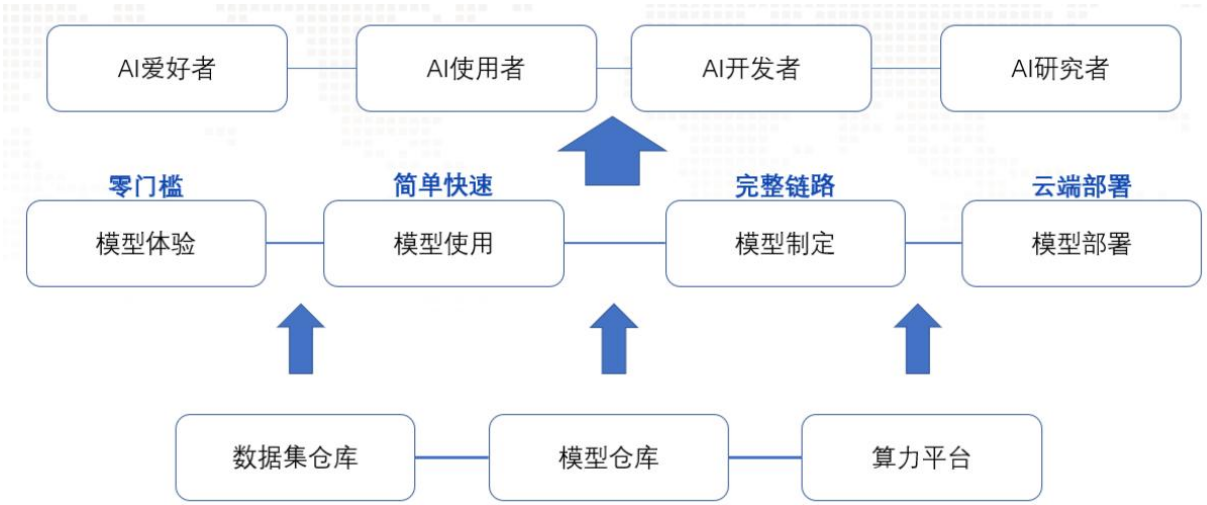
大模型厂商计费模式与服务形态的转型

国信证券 GUOSEN SECURITIES

面对 AI 定价方式从 Tokens 向 PTU / 长期承诺的转变，上游的模型厂商的收费方式与服务形态也发生了转变。

- **从API调用到企业级订阅：**API调用按Token用量向开发者及中小企业收费，模式灵活但成本不可预测。当前大模型厂商更倾向推出面向大型企业的“企业级订阅”或“专用实例”服务，客户签订年付或多年合同，可获固定调用额度、更高SLA、专属算力资源及优先技术支持，该模式为企业提供成本确定性，也为厂商锁定长期收入。
- **从公有云到私有化部署：**金融、医疗、能源等行业客户对数据安全、隐私防护与低延迟性有着极高要求，不愿将敏感数据上传至公有云。大模型厂商提出大模型私有化部署方案：客户购置硬件服务器后，可将大模型直接部署于自有私有云或本地数据中心。收费采用“一次性授权费（涵盖模型许可与部署服务）+年度维护费（覆盖模型迭代与技术支持）”模式，彻底摆脱Token计费限制，为客户赋予最高等级的控制权与安全性。

图11：MaaS产业结构



资料来源：阿里云官网，国信经济研究所整理

- **从单一模型到平台服务（MaaS）：**大模型厂商正从主要提供单一的基础模型API，加速演进为模型即服务（Model-as-a-Service, MaaS）的全新形态。大模型厂商不再局限于输出模型本身，而是围绕客户的全流程AI需求，搭建起一套覆盖数据清洗、特征工程等数据处理环节，适配多场景的模型微调服务，灵活高效的部署运维方案，以及实时动态的监控管理体系等在内的完整工具与服务矩阵。客户无需再投入大量成本搭建独立的AI基础设施，只需像订阅云服务一样，按需订阅整个AI能力平台，就能快速启动AI应用开发，大大降低了AI应用的技术门槛与试错成本。

目录

风险提示

国信证券 GUOSEN SECURITIES

- AI应用落地不及预期、市场需求不及预期、行业竞争加剧、宏观经济波动、新技术研发不及预等。

免责声明

国信证券 GUOSEN SECURITIES

国信证券投资评级

投资评级标准	类别	级别	说明
报告中投资建议所涉及的评级（如有）分为股票评级和行业评级（另有说明的除外）。评级标准为报告发布日后6到12个月内的相对市场表现，也即报告发布日后的6到12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。A股市场以沪深300指数（000300.SH）作为基准；新三板市场以三板成指（899001.CSI）为基准；香港市场以恒生指数(HSI.HI)作为基准；美国市场以标普500指数(SPX.GI)或纳斯达克指数（IXIC.GI）为基准。	股票投资评级	优于大市	股价表现优于市场代表性指数10%以上
		中性	股价表现介于市场代表性指数±10%之间
		弱于大市	股价表现弱于市场代表性指数10%以上
		无评级	股价与市场代表性指数相比无明确观点
	行业投资评级	优于大市	行业指数表现优于市场代表性指数10%以上
		中性	行业指数表现介于市场代表性指数±10%之间
		弱于大市	行业指数表现弱于市场代表性指数10%以上

分析师承诺

作者保证报告所采用的数据均来自合规渠道；分析逻辑基于作者的职业理解，通过合理判断并得出结论，力求独立、客观、公正，结论不受任何第三方的授意或影响；作者在过去、现在或未来未就其研究报告所提供的具体建议或所表述的意见直接或间接收取任何报酬，特此声明。

重要声明

本报告由国信证券股份有限公司（已具备中国证监会许可的证券投资咨询业务资格）制作；报告版权归国信证券股份有限公司（以下简称“我公司”）所有。本报告仅供我公司客户使用，本公司不会因接收人收到本报告而视其为客户。未经书面许可，任何机构和个人不得以任何形式使用、复制或传播。任何有关本报告的摘要或节选都不代表本报告正式完整的观点，一切须以我公司向客户发布的本报告完整版本为准。

本报告基于已公开的资料或信息撰写，但我公司不保证该资料及信息的完整性、准确性。本报告所载的信息、资料、建议及推测仅反映我公司于本报告公开发布当日的判断，在不同时期，我公司可能撰写并发布与本报告所载资料、建议及推测不一致的报告。我公司不保证本报告所含信息及资料处于最新状态；我公司可能随时补充、更新和修订有关信息及资料，投资者应当自行关注相关更新和修订内容。我公司或关联机构可能会持有本报告中所提到的公司所发行的证券并进行交易，还可能为这些公司提供或争取提供投资银行、财务顾问或金融产品等相关服务。本公司的资产管理部门、自营部门以及其他投资业务部门可能独立做出与本报告中意见或建议不一致的投资决策。

本报告仅供参考之用，不构成出售或购买证券或其他投资标的要约或邀请。在任何情况下，本报告中的信息和意见均不构成对任何个人的投资建议。任何形式的分享证券投资收益或者分担证券投资损失的书面或口头承诺均为无效。投资者应结合自己的投资目标和财务状况自行判断是否采用本报告所载内容和信息并自行承担风险，我公司及雇员对投资者使用本报告及其内容而造成的一切后果不承担任何法律责任。

证券投资咨询业务的说明

本公司具备中国证监会核准的证券投资咨询业务资格。证券投资咨询，是指从事证券投资咨询业务的机构及其投资咨询人员以下列形式为证券投资人或者客户提供证券投资分析、预测或者建议等直接或者间接有偿咨询服务的活动：接受投资人或者客户委托，提供证券投资咨询服务；举办有关证券投资咨询的讲座、报告会、分析会等；在报刊上发表证券投资咨询的文章、评论、报告，以及通过电台、电视台等公众传播媒体提供证券投资咨询服务；通过电话、传真、电脑网络等电信设备系统，提供证券投资咨询服务；中国证监会认定的其他形式。

发布证券研究报告是证券投资咨询业务的一种基本形式，指证券公司、证券投资咨询机构对证券及证券相关产品的价值、市场走势或者相关影响因素进行分析，形成证券估值、投资评级等投资分析意见，制作证券研究报告，并向客户发布的行为。

国信证券
GUOSEN SECURITIES

国信证券经济研究所

深圳

深圳市福田区福华一路125号国信金融大厦36层

邮编：518046　　总机：0755-82130833

上海

上海浦东民生路1199弄证大五道口广场1号楼12楼

邮编：200135

北京

北京西城区金融大街兴盛街6号国信证券9层

邮编：100032