

硅谷坐标 x FundaAI周默：美股科技公司Q2财报深度拆解：资本开支增速重新超过AI收入 – 图文解说

一张关键画面对应一段完整解说，按原视频时间推进。

文章摘要

本期围绕《硅谷坐标 x FundaAI周默：美股科技公司Q2财报深度拆解：资本开支增速重新超过AI收入》展开，按时间顺序讨论开场预告：回调、开源模型与 GCP、行情复盘：仓位、CAPEX 与流动性危机、微软财报：Copilot 加速与 Azure 需求、Copilot 的指标：席位、活跃率与 Consumption。以下内容保留完整问答、例子、数字、画面信息和限定条件，适合先读这段摘要建立框架，再结合每个关键帧逐段阅读。

01. 开场预告：回调、开源模型与 GCP（00:02–01:06）



场景 1

本节主旨 开场预告：回调、开源模型与 GCP

完整内容

周默：过去一个月可能是我们有股票以来，人类历史上回调最猛烈的一次。**开源模型的定价本质上是 CSP 定价，不是开源模型定价，所以你有了定价的能力。**这是微软在 AI 里面的逻辑。微软是这几家 CSP 和 Meta 里面唯一一个提到明年现金流为正的公司，其他几家的 CAPEX 都是负的。这个季度 GCP 的加速增长跟 Gemini 没有太大的关系，更多还是跟 **TPU 和 GPU 的销售**有关系。第一个原因是谷歌的 position 受到了不少竞争的影响，因为谷歌是一家对数字供应商筛选和数据购买都要求非常严格的公司。第二个原因是算力也是瓶颈，谷歌 Gemini 在二季度所训练的算力也就只有 **128 的一半**。你的算力也少，数据也少，就造成了 Gemini 在二季度相对来说比较落后。

画面说明

冷开场快速切换网页、数据中心、图表和谷歌相关画面，随后回到男嘉宾近景；画面下方保留节目水印与字幕。

02. 行情复盘：仓位、CAPEX 与流动性危机（01:11-11:14）



场景 2

本节主旨 行情复盘：仓位、CAPEX 与流动性危机

完整内容

曹卿云：欢迎周末返场《硅谷坐标》。三个月前财报季刚结束的时候，**周默**在这里留下了几个判断，今天看看财报季结束之后观点有没有更新。先聊最近一个月的剧烈行情：**7 月**半导体板块蒸发了超过 **1 万亿美元**市值，四家大厂的年度资本开支指引继续上调；财报之后市场反应却很分裂，微软和亚马逊被大幅奖励，Meta 下跌，谷歌先跌后修复。同样是花钱，待遇天差地别。你觉得市场叙事，特别是整体 AI 资本开支与回报的叙事，有什么变化？

周默：过去一个月可能是我们有股票以来回调最猛烈的一次。当然对 Mega Seven 来说其实没有回调，现在它们都在新高；但大部分半导体股票几乎在一个月时间跌了一半，甚至一半以上。即使回顾当年的 **dot-com 泡沫**，也没有经历过这么剧烈的回调。如果按回调幅度、速度以及参与人的感受来判断，可能比当年的 dot-com 泡沫有过之而不及。

这里有几个比较大的原因。第一，现在有了 AI 以后，所有人的学习效应和学习速度都非常快：理解利好很快，理解利空也很快，一个小时就能做完对一家公司的基本研究。第二，美股的传播速度也比一年前快很多。Twitter 和 Substack 里面各种 influencer 的影响力，越来越取代投行 Research Shop 的影响力。一个 rumor 到大家 debate，再到股价反应，速度远远比以前快。

如果复盘过去一个多月的整体下跌，5 月开始美国机构上了很高的仓位。此前大家半导体平均配 10%—15%，但 5 月几家很有影响力、也带有 influencer 属性的基金提到半导体仓位可以加到 30%—35%，所以整个 5 月美国机构上仓位的速度可能是历史最快的一个月。6 月已经有很多 token maxing 的讨论，从机构角度开始加一些短端仓位，机构净仓位已经开始下降；但散户半导体仓位应该在历史最高点，部分领域还有很高杠杆。只要出现一点下跌，仓位稍微下降，马上会引来很多关于基本面的讨论。

第一个讨论就是 CAPEX 和 ROI。我们上次聊到，今年 1 月到 5 月，半导体仓位大幅上涨有两个原因。第一，整体全市场 AI ARR 增长快于 CAPEX 增长，ARR 与 CAPEX 的差距在缩小。Anthropic 相比去年底差不多增长了 8 倍，OpenAI 相比去年底也增长了一倍多，但 CAPEX 没有同等幅度增长，所以绝对值差距在缩小，整体 AI ROI 在上半年变好。第二，每吉瓦数据中心的 revenue ARR 越来越高。去年底大家认为一吉瓦数据中心租出去一年大约带来 100 亿美元营收，现在已经到 300 亿美元；同样算力有了更高价值，ARR 与 CAPEX 的差距变小，ROI 看起来变好。

但这个季度出现了反转。上个季度我们特别讨论过，为什么几家大厂 CSP 的 CAPEX 这么强，当时一个很大的原因是存储以及半导体上下游都在涨价。经过 3 个月，涨价幅度过多了。以一吉瓦数据中心成本看，去年底差不多 350 亿美元；到明年底，考虑 B300 或 Vera Rubin，一吉瓦成本差不多要到 550 亿—600 亿美元，单吉瓦成本要涨 50%。从二季度开始供应链涨价并向下半年传导，二、三季度涨价尤其猛，CAPEX 增长又快过 ARR。CAPEX 比 AI ARR 增长快，很大原因是供应链通胀，于是许多微观公司出现长期 ROI 担忧。

典型例子是 Meta。供应链涨价后，Meta 明年 CAPEX 差不多要到 2200 亿美元，相当于明年收入的 70%、明年利润的两倍。原来大家讨论 Meta 是会把现金流花光，把所有利润拿去投 CAPEX；到了 2027 年，问题不只是把所有利润投进去，而是 CAPEX 已经是利润两倍、甚至接近收入。Meta 要回答的是：这么长、这么大的 CAPEX，能不能再造一个 Meta，而不只是把现金流花光。

7 月 1 日左右，Meta 传出要对外租 GPU、租算力，半导体和 Meta 股价反馈非常剧烈。Satya 在过去两个月一直是开源模型最大支持者，讲了开源模型对 CSP 和 ROI 的帮助，于是市场又想微软是不是也可能砍 CAPEX。随后出现 Kimi K3；它对美国投资人的影响比年初 DeepSeek V4 以及刚发布

的 V4 Flash、Pro 更大，因为大家原来认为企业模型与开源模型差距还有 **6 个月**，K3 把差距缩到 **3—4 个月**，叠加 CAPEX 讨论后，市场更担心开源模型冲击 CAPEX 格局。

再往后两周，出现亚洲无法提高 LAN 价格的舆论，包括手机厂商、BAT 拒绝 LAN 报价；北美 CSP 接受报价且议价能力很好。亚洲舆论通过 Twitter KOL 迅速传到美国，恰好又遇到美国一家知名基金降杠杆、爆仓。该基金 growth 仓位很大，可能影响覆盖全美股、数百亿美元的半导体仓位。**7 月底** 半导体经历最大一轮下跌，那一周可能跌了 **30%—40%**。

所以回头看，这一轮下跌先由仓位 trigger 引起，然后是 CAPEX concern，接着担心大厂 CAPEX，出现**存储**提价，再出现开源模型，最后以流动性危机收尾。

曹卿云：然后以流动性危机收尾。接下来把四个科技大厂的财报过一遍，首先看微软。上次你对微软最谨慎，因为 Office 没有加速、Copilot 原生集成优势削弱，而且大公司的迭代速度变慢。

画面说明

场景从主持人开场近景开始，穿插男嘉宾近景、女主持人画面、数据中心/半导体 B-roll 和多张带标题的财务报表；选中的图表呈现 Meta CAPEX、AI ARR、数据中心成本等讨论所对应的框架。

03. 微软财报：Copilot 加速与 Azure 需求（11:15–20:15）



场景 3

本节主旨 微软财报：Copilot 加速与 Azure 需求

完整内容

周默：这次微软成为整个大票里面逻辑反转最猛烈的一家。季度前我们转变了对微软的态度，在报告里特别写到：这是过去一年唯一一次看到 **Copilot 加速增长**、并且在大型企业里加速渗透的季度。

对于 Azure 来讲，这个季度变化也比较明显。Copilot 对微软趋势影响很大：微软股价最好的时候大约在两年多前，正好是 Copilot 最强的时候。当时主要炒两个股票，一个是英伟达代表的 GPU，一个是微软代表的 Copilot；Copilot 当时是全世界最大的 AI 应用场景。之后经历了 Copilot 迭代慢、Sora 和 OpenAI 有更原生模型，以及合规和权限处理拖慢开发迭代等问题。

这次变化首先来自客户购买方式。以前客户先做 **10—20 个席位**的 trial，再慢慢向其他员工扩展；今年上半年、特别是二季度开始，出现很强的 FOMO。几个非常大的客户，包括一家美国金融客户，已经按 **10 万个 seat** 规模采购 Copilot，而此前只是几千 seat 的 trial，这在过去两年从没见过。

Copilot 团队也做了很多变化：consumer Copilot 和 commercial Copilot 合并，更容易管理；Copilot、Microsoft Co-Work 以及原来类似 Open Cloud 的产品开始协同，越来越往 **Copilot Super App** 演进。Copilot 也开始做 Honey，能力仍比不上 Claude，但这季度完成后没有那么差了。越来越多底层模型换成 coding 模型。coding 模型支持更长**推理链**，执行前能做更多检查和推敲，降低复杂任务出错率。

从二季度开始，Copilot 可以独立做 App，不像以前那么难用，Excel 出错率也变小。从产品变化看，Copilot 比上季度明显；但从前沿创新和更原生的 AI 组织能力看，仍比不上 Anthropic，也比不上 OpenAI。恰好这个季度是 Copilot 迭代最快、最容易补差距的季度，所以我们很难判断变化能否延续到后两个季度，但本季度和下季度变化确实明显。过去一年中，我们第一次觉得 Copilot 在本季度会超过预期。

第二件事是 Azure 的变化。过去 OpenAI 使用 Azure 训练时有一个 super cluster，按之前协议不能算作 Azure 收入；随着 OpenAI 在 Azure 有了更多其他 cluster，这些新的训练 cluster 可以进入 Azure 收入。从今年年初开始，OpenAI 有越来越多训练收入 book 到 Azure 增长里。

二季度，Azure 在 Wisconsin 的数据中心释放了更多 capacity，恰逢算力最紧缺、最夸张的季度。Anthropic 租 SpaceX 数据中心，差不多一吉瓦一年要到 **300 亿美元以上营收**，远超过 Azure 之前租数据中心的价格。4—6 月开始，续约 GPU 与新签 GPU 价格都涨价，幅度达到 high single digit 和 low teens；存量合同不续约可能不涨，新 CAPEX 带来的客户则要涨。CPU 也涨价，可能通过收紧折扣、调整合同、让客户换新一代 instance 来实现。

同时 OpenAI 二季度收入比一季度加速，Codex 的增长基本从 **5 月**开始；OpenAI 在二季度**推理**加速，带来了 Azure 更快的加速。

除了基本面，微软业绩会还展示了一个逻辑变化。Meta 和微软同日开业绩会，Meta 先公布后半导体开始下跌；Meta 收入指引略 miss，CAPEX 本季度稍低但后两个季度更高，明年 CAPEX 更高。微软业绩会上，Satya 做了三组 ROI 对比：自己训练模型的 ROI、套别人模型再卖的 ROI、自己有模型并把模型卖出去的 ROI；微软既和 OpenAI 合作，也和开源公司合作。最引人注目的是开源模型合作：他说 ROI 差不多是 **30**，大约 **三年回本**。

这第一次公开以业绩形式告诉大家，微软为什么做开源模型，以及为什么开源模型会带来更好的 ROI。微软原来的担忧是，大部分算力租给 OpenAI，而 OpenAI 议价能力巨大，margin 和 ROI 以后提不上去；如果把模型 token 分散到更多开源模型，议价能力会变强。**开源模型的定价本质上是 CSP 定价，不是开源模型定价，所以微软获得定价能力。**

特别有意思的是，微软是这几家 CSP 和 Meta 里面唯一提到明年现金流为正的公司，其他几家的 CAPEX 都是负的。微软可以在现金流为正的情况下不断提高 CAPEX，而且增加的 CAPEX 还能带来收入加速。今年数据中心的一个瓶颈是融资供给；现金流为正会让微软在融资市场上拥有更强议价能力，拿到更多资金，甚至不完全依赖融资，而是依赖自有资金做 CAPEX。因此，positive CAPEX 也能让微软在融资市场拿到更大份额。微软这次还给了罕见的透明披露：商业 RPO 增长 **84%**。

画面说明

主要画面是男嘉宾讲解，关键帧为 Azure 与 Copilot 的财报对比表；中间出现 Copilot Super App、模型切换、OpenAI/Azure cluster 与 ROI 逻辑相关的屏幕图表。

04. Copilot 的指标：席位、活跃率与 Consumption (20:16–25:38)



场景 4

本节主旨 Copilot 的指标：席位、活跃率与 Consumption

完整内容

曹卿云：商业 RPO 达到 6000 多亿美元，剔除 OpenAI 后增长也有 25%，而且增量全部来自 Frontier Lab 以外的客户。这组披露是否能解决大家对客户集中度和低质量收入的质疑？

周默：没有解决。大部分 RPO 还是来自 OpenAI，所以增长主要还是模型公司带来的增长。不过客户集中度出现了一个好的趋势：微软开始寻求 OpenAI 以外更多算力需求方，并且能够掌握算力定价能力。之前如果做后测开源模型，开源模型定价就是微软做的。

这季度 Azure 毛利率已经到了 **85%**。OpenAI 的毛利率肯定在 **50%** 或略高；原来做 GPU 租赁，毛利率差不多 **47%**。整个环节里模型公司赚得比 CSP 多，但如果后测开源模型，就可以把模型公司那部分利润拿到 CSP 环节。因此，无论做 token factory 还是开源模型，margin 都肯定比原来与 OpenAI 合作更高。微软没有完全回答客户 diversity，但回答了 margin 和 ROI。

曹卿云：Copilot 付费席位从这一季度的 **2000 万** 升到超过 **3000 万**，新增环比也翻倍；相对于 Microsoft 365 数亿总用户，渗透率仍不高。更看重什么指标：付费席位、实际活跃率，还是 consumption？

周默：对整个 Copilot 体系，过去最重要的指标应该是激活和活跃率。大量企业买了 Copilot，但激活率和活跃率都不高；很多企业计算活跃率时只能看周活、月活，不能看日活。活跃率代表产品体验，也可以用 NPS 看产品体验。

这个季度明显的变化是，过去用 Copilot 的 PPT 独立做 PPT 几乎不可用，只能用 Codex 或 Claude；现在 PPT 加入底层 coding 模型以后，已经能按指令做出能用的 PPT，体验端二季度变化明显。所以这些指标都改善了。数字改善之后，下一步是大企业能不能做 expansion：原来只给 **5%** 的用户买 seat，如果这 **5%** 用得好，能不能扩到 **30% 或 50%**。

本季度看到几家公司一次性全年购买超过 **10 万个 seat**，过去从没见过大型跨国集团从几千个试用 base 变成全部使用的大 base。从现在往后，更要关注大型企业里面 Copilot 的渗透率。如果越来越多公司从 **5%—10%** 变成 **30%—50%**，Copilot 就到 next stage。相对于整个 Office 365，Copilot 现在仍只有约 **5%** 渗透率：一方面没有覆盖全部 Office 客户，另一方面最大的客户 Copilot 渗透率也不高。愿意加预算、愿意吃螃蟹的大客户，如果渗透率上升，就是 Copilot 渗透率上升的前沿指标。

另一个有意思的点是，Copilot 开始更多做 consumption 营收。Copilot Studio 上季度已经讨论过，很多功能现在也集成进 Copilot。加了 coding 模型底座后，可以做更多 AI-native App。虽然这些 App 仍包在 Copilot 价格里面，但触发的 token 可以按 **Consumption** 计费：先定一个限额，超过限额就按 consumption 收费。对于 AI-native 企业，单个 seat 的 Copilot 价格可能超过 **30 美元**，多出来的就是 Consumption 定价。

画面说明

画面以男嘉宾近景为主，期间切换主持人镜头；底部字幕和节目水印持续存在，没有额外可读图表，选帧保留了完整讲解构图。

05. Azure 加速的三层来源 (25:40-30:51)

⑦ Azure 增速>40% 五层瀑布

三层 OpenAI + 涨价 + 容量 → Q2 40%，下季指引 45%

OpenAI 训练	OpenAI 推理	OpenAI API	Capacity+Price
supercluster Q2 起算	Codex 5月+ 推理加速	Anthropic + 开源分发	Wisconsin 300-400MW GPU+10% CPU隐性

对客：长协续约涨价 + 新客户涨价
需求超供给 → 下季指引 45%

硅谷坐标 周默运场 深度概念图

然后第二个点的话呢 是你也可以看到OpenAI 本身在二季度

SILICON VALLEY
VECTOR
硅谷坐标

美股财报 | 资本开支 VS ROI之争 | 微软财报 | 亚马逊财报 | Meta财报 | 谷歌财报 | 明年资本开支 | 云厂商模型格局 | 英伟达财报展望

场景 5

本节主旨 Azure 加速的三层来源

完整内容

曹卿云：这次财报 Azure 同比增长 **43%**，超过自身指引，下季度还指引 **45%**继续加速。一个原因是 CPU/GPU 效率提升、新产能提前上线并迅速变现。这个加速是需求进一步加速，还是此前被压抑的需求随产能释放而确认？

周默：Azure 加速有好几个 driver，最大的还是 OpenAI 给 Azure 产生的贡献，分成三个 segment。第一，OpenAI 在微软训练产生的收入；第二，OpenAI 在微软**推理**产生的收入；第三，微软售卖 OpenAI API 带来的营收。这三点在二季度都有比较大的边际变化。

第一，OpenAI 在微软的训练收入在今年年初之前不确认，是因为 OpenAI 和微软最大的 Supercluster 里面的收入不算微软云收入。随着 OpenAI 在微软有了更多 cluster，更多不在这个 Supercluster 里的训练就可以确认进 Azure 云收入，所以微软有了更多来自 OpenAI 的训练收入。

第二，OpenAI 二季度 ARR 相比一季度加速，Codex 的反馈从 5 月开始变好。OpenAI CEO 此前讲 7 月 ARR 增速非常好，环比超过 20 个点，净新增超过此前三个季度。所以 OpenAI 的推理收入在二季度边际也变好。

第三，微软售卖 OpenAI API 与 OpenAI 推理相关；微软也可以售卖与 Codex 相关的 API，以及 OpenAI 其他 API。

除了 OpenAI，Azure 增速无非是 价格 × capacity。对云最重要的变化就是 GPU 价格乘以 capacity。微软以前经常每季度放不出 capacity：可能被 OpenAI 拿去训练，而 Supercluster 训练收入又不算微软收入，也可能投入其他产品。这个季度微软在 Wisconsin 的数据中心放出了约 300—400 兆瓦 capacity，而且这批 capacity 基本都能用来 generate 营收。

第二个点是价格。真正 GPU 大幅涨价从今年二三月份开始，GPU spot pricing、on-demand pricing 和长协价格都在涨。最紧缺的时间点大约是六七月份。Anthropic 甚至愿意接近 spot pricing 的价格签 SpaceX 300—500 兆瓦数据中心。我们调研下来，微软最近一个季度新签客户的 GPU 价格比之前涨了约 10%；存量客户没到续约不能涨，续约和新签客户都要涨价。

微软每年数据中心的吉瓦数基本翻倍，所以今年涨价在新客户里都能计算，到年底大约一半客户的 GPU 可以涨价。除了 GPU capacity 和价格都放出来，CPU 也开始涨价，但 CPU 涨价更隐性：收紧折扣、续约时推动客户签新的 instance 等。于是这个季度微软的 OpenAI 收入、GPU 收入和 CPU 收入都变好。

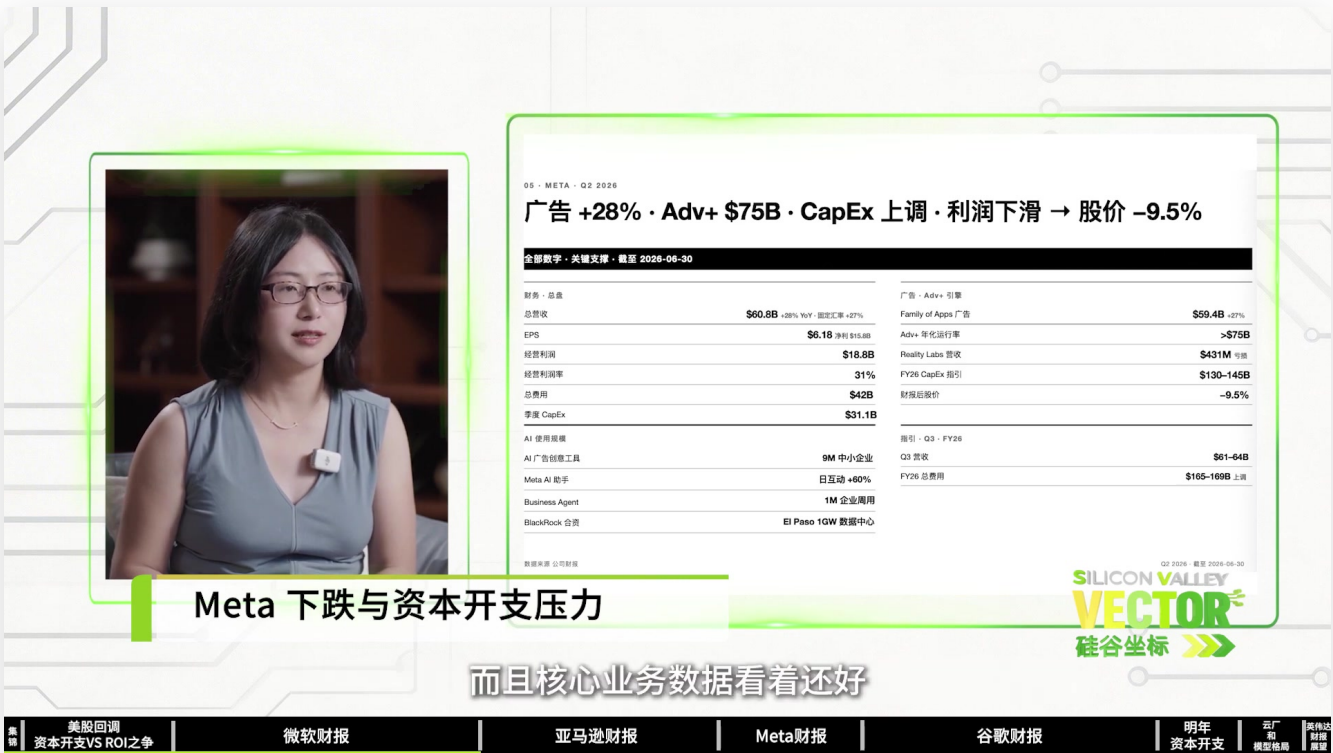
微软从一季度开始售卖 Anthropic 模型，二季度开始售卖更多开源模型。Anthropic 模型在微软 API 售卖里应该占 10%—15%，相比一季度，光卖这类模型也可能贡献营收增长。

曹卿云：好的，接下来看看 Meta。Meta 是这几个公司里面唯一一个……

画面说明

选中画面展示“Azure 加速、三层来源”的结构化图表：训练、推理和 API 三列并列，旁边嵌入男嘉宾画面；其余镜头为男嘉宾近景和节目字幕。

06. Meta 财报：下跌与 CAPEX 压力（30:52-33:14）



场景 6

本节主旨 Meta 财报：下跌与 CAPEX 压力

完整内容

曹卿云：Meta 是这几家公司里唯一一个下跌的大票。它的 CAPEX 绝对额也是四家里最少的，核心业务数据还好：广告收入涨 27%，曝光量涨 14%，单价涨 12%。花得最少、业务也不差，反而下跌。另一个细节是本季度有 24 亿美元诉讼计提和 12 亿美元裁员费用；剔除后经营利润率仍在增长，但市场没有穿透一次性费用，究竟在惩罚什么？

周默：先纠正一个点，Meta 的 CAPEX 不是最少。绝对值是最少，但 CSP 大量 CAPEX 要外租；还有原来的 lexical、CPU CAPEX。Meta 里与云、GPU、模型相关 CAPEX 的纯度和含量最高，所以按 GPU 和 AI 相关 CAPEX，它肯定不是最少。

这季度市场反应的一个原因，是我们之前提到的 **2027 年 Meta CAPEX 太大**，相对于收入体量太大。Meta 在这几家公司里面营收最少，明年大约 **3000 亿—3200 亿美元**；**2200 亿美元**CAPEX 差不多是收入的 **70%**。Meta 不只是要用这些 CAPEX 做 **10%—20% 的营收增量**，而是要回答：**占收入 70% 的 CAPEX，能不能再做出 70%—100% 的 Meta 营收？**

Meta 的营收基本分为两部分：已有广告营收，以及以后可能产生的 AI 营收。对本季度而言，大家对 AI 营收有较高预期。**投了这么多 CAPEX 以后，扎克伯格需要讲清楚这些 CAPEX 怎么产生营收：无论是卖模型 token，还是去做数据中心，都要讲出一条清晰路径。**

画面说明

关键帧是女主持人与“**Meta 下跌与资本开支压力**”的财报表格同框，表格列出广告、CAPEX 与利润率等对比；后续切换到男嘉宾近景。

07. Meta 为什么暂不外租算力（33:15–38:37）

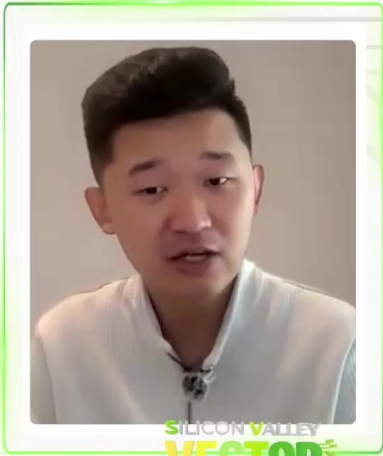
⑩ Meta 外租算力 阻力重重

卖 Token 毛利 > 卖算力毛利 → 倾向卖 Intelligence

卖 H 卡 (推理)	卖 B 卡 (训练)	卖 Intelligence
SLA 缺失 不如 neocloud	卡型 谈不拢 买方 也要 B 卡	毛利 更高 主推 方向

时间表: 2-3 季谈判 → 今年底签约 → 明年 Q1 见营收

硅谷坐标 周默运场 深度概念图



美股回调
资本开支 VS ROI 之争

微软财报

亚马逊财报

Meta 财报

谷歌财报

明年
资本开支

云厂商
模型格局

英伟达
财报展望

场景 7

本节主旨 Meta 为什么暂不外租算力

完整内容

曹卿云：二季度有个戏剧性插曲：新闻报道 Meta 准备外租算力，消息一出股价暴涨；随后公司辟谣，股价回落。市场用涨跌投票，希望 Meta 卖算力，但管理层拒绝了。扎克伯格的官方说法是，Meta 收到的算力报价显著高于自己的采购成本，但他判断卖 intelligence 的毛利会持续高于卖 compute，所以不卖。你怎么看？Meta 未来会外租算力吗？

周默：Meta 这次业绩会不是在讲“不卖”，而是在讲卖算力之外其他 CAPEX 的合理性。扎克伯格的故事是，随着 Meta 更多模型变好，去卖自己的 token 后会有比卖 compute 更高的毛利率，所以现在投入以后会有更多毛利率。

如果现在有很好的算力报价，而且报价可以满足、算力可以卖，Meta 也可能卖。但我们业绩前做了不少调研，感受是算力没有那么容易卖，主要有两个原因。

第一，做云服务有一个专业指标叫 SLA。SLA 不只是衡量出错概率，通常有“三个九”或“四个九”的概念：出问题的概率小于 1 减去 99.9% 或 99.99%。如果出了问题，服务商要按原价甚至超过原价赔偿。这是云厂商给客户的 guarantee，表示对 host 这一套服务器很有信心。对 CSP 来说，这是最简单的指标，但 Meta 以前没做过，因此没有 SLA 级别的服务能力。它 host 数据中心的能力不只比 CSP 差，也比 Neocloud、甚至 SpaceX 差，所以要从 AWS 等地方挖专业人士过来。

第二，Meta 要衡量租什么卡。Meta 特别之处是大部分算力用来训练，推理场景没有那么多：一个是推荐算法迭代，但推荐并不特别吃算力；另一个是 Meta 各种场景调用内部 token，使用量也比训练小很多。H 卡做推理性价比很高，相比 B 卡不差很多；训练则要用最好的卡，以获得最好的 ROI、训练更大的模型。

因此 Meta 想把 H 卡租出去，不想把 B 卡租出去；但算力购买方如 Anthropic 也想租 B 卡，不想只租 H 卡。买方不会接受只租 100 万张 H 卡、一张 B 卡都没有，而会要求除了大量 H 卡，还租几十万张 B 卡。谈判不会顺利，中间会有迂回，也可能谈不成；而 Meta 的 SLA 能力目前也不具备。

我们觉得 Meta 最早产生 AI 算力租赁，可能要到今年底或明年初，不会那么快。合同签完后，Meta 可能在年底讲算力报价开始对明年收入有明显贡献，但不会是本季度。市场却希望本季度就看到 AI revenue 路径。我们的测算是，这些 CAPEX 对应明年毛利率可能有 6—7 个点负面影响，后年还会有影响；今年 Meta 增长还可以，但明年利润增长可能只有 low single digit。这反映了 CAPEX 对利润的影响，所以 Meta 更急切要讲清楚 AI revenue 的路径。

后面每个季度，大家都会非常关注 Meta 的 AI revenue 是怎么产生的，以及利润率以后怎么变化。

画面说明

画面以“Meta 外租算力方案”表格为主，列出模型、算力与 intelligence/compute 的比较，男嘉宾画面嵌在表格一侧；穿插男嘉宾近景与字幕。

08. Meta 广告 AI 化与 Business Agent (38:38–46:06)

⑪ 广告自动化四阶 → Business Agent

Adv+ 年化 750亿 = 全广告 1/3; 点击+8% 转化+15%		
四阶 演进	渗透 路径	意义
人工 → Advantage+ Business Agent → 主动Agent	SMB 2025 中型 2026 → 大企业 2027	像当年 Advantage+ 一年后 或成 强制产品

硅谷坐标 周默返场 深度概念图

我觉得这个季度 Meta比较明显的变化是

美股回调
资本开支 VS ROI之争

微软财报

亚马逊财报

Meta财报

谷歌财报

明年
资本开支

云广
和
模型格局

英伟达
财报
展望

场景 8

本节主旨 Meta 广告 AI 化与 Business Agent

完整内容

曹卿云：上次你讲 Meta 广告加速主要来自产品形态和库存，单次推荐算法升级对整体 revenue 贡献不到一个点。这次 Meta 财报披露，部分 AI 模型带来 Facebook 广告点击约 **8%**增长、转化约 **15%**；Advantage+ 全自动投放年化运营率超过 **750 亿美元**，相当于三分之一广告收入已经流进 AI 自动化系统。你看到什么新变化？

周默：这个季度 Meta 明显的变化，是把原来一套做 AI 的产品 Business AI 整体 rebranding，现在叫 **Business Agent**，把大部分 AI 功能，包括新增 AI 广告功能，都放进去。

这个产品未来的价值，可能相当于 Meta 当年推动 Advantage+ 的变化。Advantage+ 支持了 Meta 2023—**2024 年**最主要的增长：先让中小企业、再让大企业拥有自动化投放和自动化做素材的能力。Business Agent 在 Advantage+ 基础上更进一步，起初也面向中小企业：你可以把所有数据接入 Business Agent，它能做数据分析、留存分析、内容分析，甚至监控竞争对手投放结果；做素材时还可以通过对话讨论如何优化素材。

这相当于从 Advantage+ 自动化投放再往前走一步，不只做优化师那一段，还做优化前的分析、决策和数据打通。**如果真正做到，就能解决大部分 SMB 的投放工作，而不只是单个优化师的工作。**这个功能会渐进变化：今年大部分时间可能推小客户、中客户，明年再慢慢推大客户。

我们与一些 Meta agency 沟通时，有些 agency 的感受像几年前第一次看到 Advantage+，认为一年后 Business Agent 可能成为 Meta 强制使用的产品。Advantage+ 最早可以主动开也可以不开，现在基本所有默认投放都用它：先对中小客户强制，再慢慢推动大客户强制。如果 Business Agent 做成，它会是广告主走向全自动广告投放的第一步。**Meta 还有巨大的分发和私有数据优势。**

曹卿云：但 Muse Spark 并不是公认最强的 frontier model。对推荐、广告、Business Agent 三个 Meta 主战场来说，模型做到足够好是否就够了？还是必须继续烧钱保住 Frontier Lab 地位？

周默：对 Meta 商业模式，做到足够好就够了。做内容、广告投放或回答基本 search 问题，还没有到非常复杂的 agent planning；而且 Meta 的数据相对 specific，做到非常强的模型，对现有业务帮助没那么大。

但如果 Meta 做成非常强的模型，又可能开拓新的增长点。Meta 有大量用户个人行为数据和行为习惯，最好的 Personal Agent 诞生地可能是 WhatsApp 或 Facebook Messenger，这是 Meta 天生的土壤；它比 OpenAI 和 Anthropic 更容易收集 consumer 端用户行为。如果有和 frontier model 一样好的模型，就有很强营收 leverage：可以收订阅价，甚至按 personal agent 收更高 consumption 费用，想象空间很强。

如果个人用户与 Meta 达到这种互动层面，更多行为、电商偏好和旅游偏好可以以脱敏的 group of data 形式提供给 Meta，不一定是敏感数据。**Meta 能拿到更多销售线索，尝试线索导向或电商导向广告，这些是 Meta 过去不太擅长的，也能帮助广告收入，但风险会更大。**

随着数据生态发展，越来越多公司能提供强化学习数据，不只是 IO 级别，也有 pre-training 级别。这会帮助 Tier 2、Tier 3 开源模型及 Meta 缩短与前沿模型差距。数据更丰富，Meta 训练也可能加速迭代。**所以此时 Meta 继续投入前沿模型的性价比比以前高；但放在整个公司，CAPEX/ROI 的讨**

论又最激烈。对管理层而言，很难放弃加大前沿模型投入，因为这可能正是追上前沿模型的最好时间点。

画面说明

选中的 Business **Agent** 图表把三类角色/流程并列，旁边嵌入男嘉宾；场景中也出现女主持人近景、Business AI/Advantage+ 字幕和节目水印。

09. 亚马逊财报：AWS 连续五季加速（46:07-50:26）

⑫ AWS 增速 36.7% 双驱动

Anthropic ARR + Blackwell 拉货 贡献几乎等量		
Anthropic	Blackwell	展望
ARR 涨 110-120% 6月底 vs 3月底	Q1 拉货 Q2 才变营收 (3月时差)	Q3 41-42% Q4 维持

意义：拉货 ≠ 营收，隔季兑现 → 逻辑单纯

硅谷坐标 周期运场 深度概念图



SILICON VALLEY
VECTOR
硅谷坐标

AWS 营收就有机会能够到41%~42%

美股回调
资本开支 VS ROI之争

微软财报

亚马逊财报

Meta财报

谷歌财报

明年
资本开支

云厂
模型格局

英伟达
财报展望

场景 9

本节主旨 亚马逊财报：AWS 连续五季加速

完整内容

曹卿云：接下来看看亚马逊财报。这次 AWS 业绩很亮眼：上个季度增速只有 28%，被质疑掉队；本季度达到 36.7%，连续第 5 个季度加速，是 18 个季度里最快的，单季环比新增收入超过 46 亿美元。

周默：这比历史最大增量还高 80%。最主要的加速来自 Anthropic。Anthropic API 营收绝大部分应该是 cloud API，也就是通过 CSP 销售；CSP API 里绝大部分通过 AWS backend 售卖，AWS 在这个 revenue model 里可以拿走 40%，所以 Anthropic API 营收能直接供应到 AWS 营收。Anthropic 二季度增长比一季度更快，6 月底 ARR 比 3 月底增长 100% 出头，体量在 AWS 里面也更大，直接带动 AWS 增长。

同时 AWS GPU 也放出更多 capacity，所以本季度 AWS 营收做到接近 37%。今年下半年和明年 AWS 营收可能更快，增长逻辑相对清晰，主要是两个 driver。第一是 Anthropic 增速：Anthropic API 差不多 70%通过 CSP 销售，CSP 销售中将近 90%通过 AWS。二季度 AWS 占比肯定有所降低，因为微软也通过 GCP 销售 token，但 AWS 仍能在商业模式中抽走 40%。Anthropic 6 月底 ARR 比 3 月底增长约 110%—120%，基数更大，所以对 AWS 二季度营收帮助明显。

第二是 Blackwell 对 AWS 的帮助也明显。AWS 从今年一季度开始加速拉货 Blackwell，最大出货量变化在一、二季度。但一季度 Blackwell shipment 不会立刻变成收入，通常要隔 3 个月，调试并放入数据中心后，到二季度确认收入。因此 Blackwell 加速从二季度开始，这也是上季度观察到 shipment 却没有反映到 AWS 营收的原因。

上季度我们判断 AWS 二季度增速能到 36%—37%，本季度是 37%，就是因为 shipment 到收入确认有一个季度时间差。Blackwell 的帮助会持续到下季度：二季度拉货比一季度更大，一、二季度的拉货都会确认到三季度收入，所以三季度 Blackwell 对营收增速的贡献会比二季度更明显。Blackwell 对 AWS 的营收贡献，和 Anthropic 对 AWS 的贡献几乎一样。

我们判断三季度 AWS 营收有机会到 41%—42%，四季度可能也是这个数字，到明年才慢慢下降一点。

画面说明

选中画面是 AWS/Anthropic/Blackwell 的三列对比表，图表中保留了季度、AWS 增速和 CAPEX 等字段，旁边嵌入周默近景。

10. AI 如何拉动传统云（50:27–55:43）

⑬ 涨价不止 GPU 更是 CPU

Graviton 每季 +2-3% 隐性涨价 (Q2明显 Q3更大)		
AI 用 CPU 三层 训练 GPU / 后训 RL Agent 调用 长时 CPU	涨价 机制 折扣减少 + 换新一代 CPU 不是商品 不会 +10%	Trainium 利用率 40-50% → 80% 进 margin 后退
意义: AWS CPU 收入 + 毛利 + ARR 齐升		

硅谷坐标 周期运场 深度概念图

对亚马逊的变化 下个季度会更明显



SILICON VALLEY
VECTOR
硅谷坐标

美股回调
资本开支 VS ROI 之争

微软财报

亚马逊财报

Meta 财报

谷歌财报

明年
资本开支

云厂
格局

英伟达
财报展望

场景 10

本节主旨 AI 如何拉动传统云

完整内容

曹卿云：亚马逊 CEO Jesse 在电话会上讲了一个重要机制：AI 在拉动传统云。我的理解是，后训练、强化学习、Agent 调用工具的计算主要跑在 CPU 上而不是 GPU 上，所以 AI 每花一块钱，都会外溢成传统云消费。如果机制成立，怎么看 AWS 的 AI 收入？是不是不能只看 GPU 和 Trainium，而要看整个 AWS consumption？

周默：这个季度我们观察到 Graviton 价格差不多涨 2%—3%，已经反映 AI workflow 对 CPU 的带动和 CPU 价格带动。但 CPU 不是商品属性特别强的产品，不能像存储一样理解涨价。它对整个云的贡献是缓慢渐进的，每季度可能有两三个点变化，不能指望供需缺口一下子带来每季度 10 个点或更

多变化；也可能有两三个点反映在利用率上。从本季度开始已经有 CPU 对亚马逊的变化，下季度会更明显，因为 CPU 涨价幅度会比二季度更高。它不是阶梯型、一下子贡献很多的 driver，而是每季度贡献一点。

曹卿云：AWS AI 收入占比只有 **15%**，看起来低于 Azure 的约三分之一。这个占比低，是否恰恰说明拉动效应更强？

周默：三家 CSP 收入结构不一样。AWS 有两个口径，一个叫 chip revenue，一个叫 AI revenue。AI revenue 更偏软件，chip revenue 更偏几张芯片，包括 training、Graviton、Nitro 等，所以实际数字会比 **15%** 高，可能在 **20% 左右**。横向比较，谷歌 AI 占 revenue 比重差不多 **45%**，微软 Azure 约 **30%**，大致是 **20%、30%、45%**。

最直接原因是原来 CSP cloud 的大小。市场电很缺，CAPEX 也很缺，在 CAPEX 相对有限时，反映的是原来的 legacy cloud 有多大。不能指望 AWS 现在 **20% 占比**以后和谷歌一样多，但 AWS 的 **20%** 以后可能到 **40%**，谷歌 **45%** 以后可能到 **60%**。在 AWS 这么大基数上，带动 **10 个点**增速已经非常不容易。

曹卿云：市场原本担心 AWS 的 AI 云收入稀释利润率，但 AWS 本季度利润率接近 **39%**，这是 Trainium 和 Graviton 的经济性开始兑现了吗？

周默：利润率反映现在的市场紧缺程度。二季度 GPU 和 Trainium 的紧缺程度都比一季度更强，定价空间更好，所以毛利率更好。如果供给上来、没有那么紧缺，毛利率会下来一点，但至少一、二、三季度可以保持。

对 AWS 来说，相对别人的 alpha 是今年一季度以前 training 利用率特别低，需要给很深折扣；从一季度开始不用给深折扣，利用率也上来。原来可能只有 **40%—50%**利用率，现在做到 **70%—80%**，相当于同样成本下有更多 capacity。这个趋势二季度也会延续，但现在纯电利用率已经很高，不能指望训练继续给 AWS 带来更多 margin 帮助。

曹卿云：亚马逊说完，最后一家是谷歌。上季度 Gemini 3.5 还是追赶中的希望，这季度谷歌推迟 Gemini 3.5 Pro，管理层罕见承认 coding 和.....

画面说明

选中的是 Graviton/Trainium 相关的芯片与传统云对比图表，画面同时保留**周默**小窗；场景后段回到男嘉宾近景并由主持人转入谷歌。

11. 谷歌财报：Gemini 迭代为何落后（00:55:43–01:03:04）

①④ Gemini 落后 = 数据 + 算力

不是组织问题 → 解决数据+算力 即可

XAI 占位 T1.5 占位位置 难以超越	数据 慢半拍 采购路径严 谨慎 慢于 其他
算力 一半 1GW = OpenAI 一半 难支撑 扩面	Gemini 4 补 年底 重新训练 更大算力 + 数据补

硅谷坐标 周默运场 深度概念图

XAI 在收购过Cursor 以后

SILICON VALLEY
VECTOR
硅谷坐标

美股回调 资本开支 VS ROI之争 | 微软财报 | 亚马逊财报 | Meta财报 | 谷歌财报 | 明年 资本开支 | 云厂商 模型格局 | 英伟达 财报展望

场景 11

本节主旨 谷歌财报：Gemini 迭代为何落后

完整内容

曹卿云：谷歌推迟 Gemini 3.5 Pro，管理层也承认 coding 和 **Agent** coding 需要提升，并把希望放在更大的 Gemini 4 基础模型上；同期 API token 处理量从每分钟约 **160 亿**涨到约 **220 亿**，云增速达到 **82%**。**模型落后和业务狂奔同时发生，这是谷歌的核心矛盾**。从 Gemini 角度看，这是正常迭代落后，还是组织和 scaling 路线有问题？

周默：本季度 GCP 加速增长跟 Gemini 没太大关系，更多跟 TPU 和 GPU 销售有关。GCP 最大增长贡献是谷歌直接向外卖 TPU，差不多卖 **10 万张卡**，带来约 **20 亿美元**营收，量非常大。

Gemini 二季度环比增速比一季度明显变慢：一季度可能超过 **50%**，二季度可能只有 **20%—30%**。这个季度增速可能只和 OpenAI、Anthropic 的月度增速差不多快。

第一个原因是谷歌的 position 受到竞争影响。原来谷歌是唯一一个 T1.5 的公司：T1 是 OpenAI 和 Anthropic，T1.5 是谷歌，再往后是其他模型公司。现在 T1.5 的公司更像 xAI；xAI 收购 Kensho 后补充了大量高质量数据，甚至比此前预想的更好。

第二个问题是谷歌合规要求严格。在数据供应商筛选和购买数据时，谷歌比其他 lab、尤其开源公司更谨慎。上季度我们讲过，谷歌去年没有怎么购买用于 post-training 的数据，也没有怎么做 post-training；今年 1 月开始购买并做 post-training，但从 5 月开始数据生态又发生很大变化。除了原来的数据供应商，还有更多、更高质量的数据提供方，而且不止一两家，来源也比纯靠人员标注的方式丰富得多。

各类 lab 能获取的高质量数据源可能比一季度多很多倍；但谷歌采买路径严格、速度偏慢，基本比其他公司买数据慢半拍，于是别人有丰富数据，Gemini 的数据反而成为最大瓶颈。

第三是算力瓶颈。谷歌一直按前沿模型方式做，但 Gemini 二季度训练算力只有 OpenAI 和 Anthropic 的一半，差不多只有 1 吉瓦用于训练。算力少、数据也少，造成 Gemini 二季度相对落后。从 coding 能力看，Gemini 可能比不过 Kimi K3、GLM 5.3 和 xAI 4.6。

以后再关注 3.5 Pro 的价值已经不高，肯定需要新的数据源和更大算力重新训练，下一代要看 Gemini 4。我不认为 3.5 Pro 能赶在 9 月或 10 月发布；Gemini 4 大概率要到年底或明年初。如果谷歌补上更多数据源并提高算力，know-how 没有特别大的差别。现在模型训练已经很工业化：不同团队找不同数据，再 merge，持续跑 data experiment 验证数据可信度和 contribution，然后归到统一模型训练方式，和一年前逻辑已有变化。

只要谷歌能获取更多数据，Gemini 4 追上只是早晚，没有特别多组织架构问题。组织问题可能是上周讨论较多的 Jeff Dean 离开，以及 Demis Hassabis 可能退居二线。DeepMind 有一批人以诺贝尔奖为目标，也有一批人以训练前沿模型为目标；现在大部分工作由后者主导，所以追求诺贝尔奖的研究者去做其他创业，不影响谷歌现在的工作。

曹卿云：Gemini 在 benchmark 暂时落后，但 API token 和 GCP 仍高速增长。对企业客户来说，是不是 Flash 的成本、延迟和稳定性比 Frontier 能力更重要？

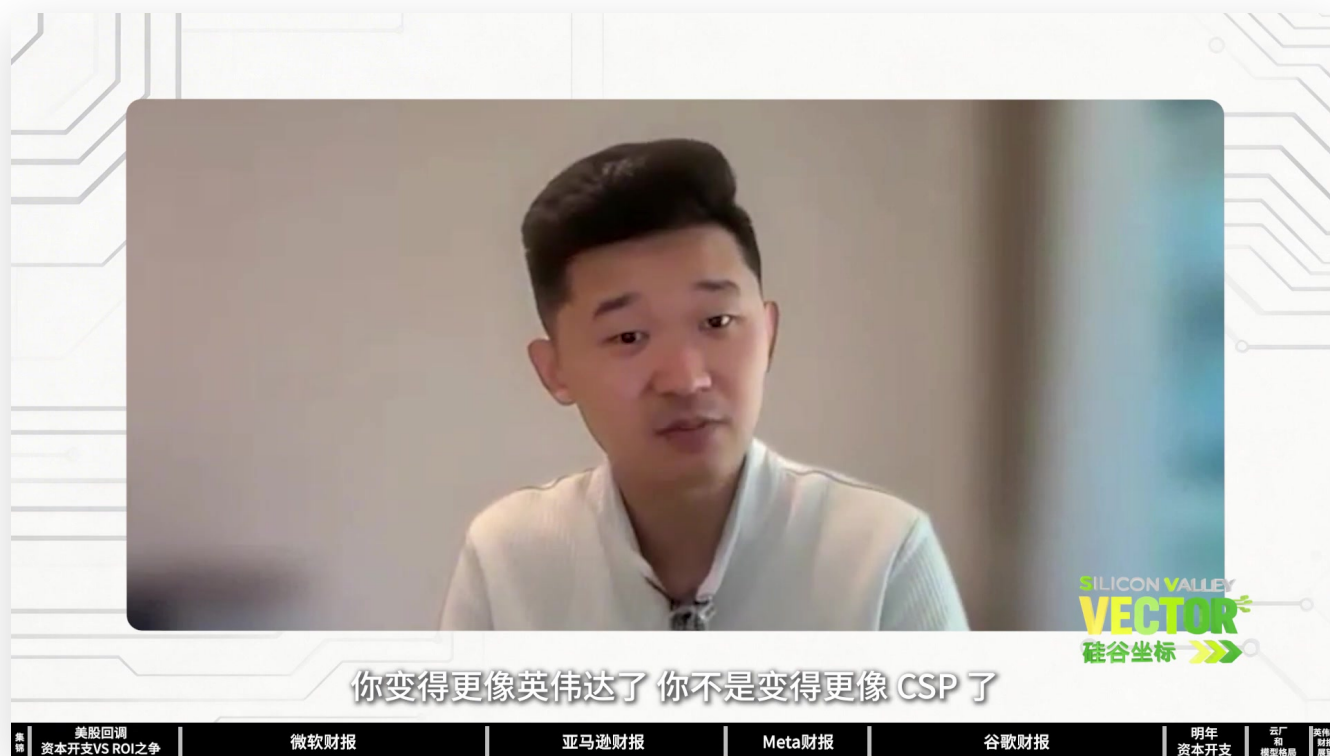
周默：不是。如果按 Flash 性价比，很难比 xAI 强，也很难和开源模型竞争。谷歌要有更多 Gemini 营收，还是要靠前沿模型和 coding 营收。最好比前沿模型慢 10%—20%，但性价比有 30%—50% 优势，这样竞争力才强；如果归入开源模型竞争，就很难比开源模型更强。

回到 GCP 贡献，这次主要是 TPU 和 GPU，Flash 对这次 GCP 加速应该没有贡献。

画面说明

关键帧为“Gemini 落后、数据与算力”四格图表，男嘉宾以小窗嵌入；场景内也有男嘉宾和女主持人近景、下三分之一标题字幕。

12. GCP 82% 增长：TPU 直销与租赁（01:03:05–01:09:20）



场景 12

本节主旨 GCP 82% 增长：TPU 直销与租赁

完整内容

曹卿云：顺着刚才的话题，这次 GCP 增长 **82%**，口径包含 Gemini、Vertex AI、TPU 租赁等不同名目，云 backlog 也达到可能 **5000 多亿美元**。**82%** 里有多少是可持续 recurring consumption，多少是硬件销售或大客户合同的一次性确认？

周默：之前讨论很多的是直销 TPU 是否是 recurring revenue。GCP 第一次直接销售芯片后，更像英伟达，不再只是 CSP，这是谷歌越来越大的变化。

但如果看 TPU shipment，明年是今年的 3 倍，明年可出售数量可能不止今年三倍，所以明年 TPU 仍会加速。到 2028 年可能无法像 2026、2027 年提供那么大的变化，TPU 增速可能比 GCP 增速慢，会拉低 GCP 增速；但 2027 年 TPU 增速肯定比 GCP 中除 TPU 外的增速更快，会把 TPU 增速

拉上去。至少明年 TPU 还能继续贡献 GCP 加速，收入贡献比今年大得多。

其他收入如 TPU 租赁、Vertex AI、Gemini，都与传统 CPU 云相似，是容易定位的 recurring consumption，没有太大疑问。**最大问题是直销 TPU：它是一次性确认 revenue，不能分到四年确认。**但 shipment 明年仍高速增长，至少明年会继续贡献增量。

曹卿云：现在 GCP backlog 里有多少是 TPU 系统销售和少数 AI 客户？多少是传统企业 workload？这对未来增速和毛利率意味着什么？

周默：TPU 直销毛利率挺高，差不多 **50%**：一张卡卖 **2 万美元**，博通成本 **1 万美元**，最后还能有 **1 万美元** 毛利率，所以比 GCP 毛利率高，也比卖 CPU/GPU cloud 的毛利率高。我们估计，从三季度往后算 **6 个季度**，直销 TPU 营收应做到 **500 亿—600 亿美元**；二季度直销 TPU 差不多 20 亿，往后六个季度做到 500 亿—600 亿，相当于 backlog 的 **10%**。现在直销 TPU 营收占 GCP 也约 **10%**。

GCP 利润率从原来 **20%** 升到 **35%**，但下季度为缓解产能短缺，要采用更多第三方算力，承受利润压力。

曹卿云：这是否说明 TPU 全站成本优势仍不足以解决供给问题？

周默：最大的原因还是没有足够电、没有足够 capacity。电不够，所以向外租 Neocloud capacity，主要租 GPU；现在没有多少新的 TPU capacity 可以给，TPU 大部分靠自己，或者靠 FreeStack 一类供应。

这反映谷歌对外销售的算力不够，需要通过 Neocloud 补可扩展性，和 TPU 成本没有直接关系。但这也是谷歌本季度业绩下跌的主要原因：谷歌 CAPEX 已经很强，本季度少一点、后面季度更高，明年 CAPEX 要到 **3000 亿美元以上**，是这些 Mega Seven 里最高的；同时又说 3000 多亿的 capacity 不够，还要去外面租 Neocloud capacity。

租别人的 capacity 不算 CAPEX，算 OPEX，会影响毛利率，实际 CAPEX 看起来更高。**所以业绩后谷歌下跌，原因是大家不确定 GCP 未来 margin，也觉得 CAPEX 看起来更高。**

但 GCP 向 SpaceX 租 capacity，价格肯定高；如果去租，肯定已经锁定用量和需求，知道比如 **500 兆瓦** 需求能满足且有利润。不给这些用户需求，他们就会跑去其他云，影响份额。所以 GCP 做这件事边际上肯定有利润，只是可能影响整个云的利润。SpaceX 定价很贵，很多按 spot pricing 和 on-demand pricing，无法按照长约 contract pricing 赚同样的 margin。

画面说明

选中画面是“GCP **82%**、TPU 直销与 Vertex”结构表，列出 TPU shipment、直销收入、毛利率与 backlog；画面旁有男嘉宾嵌窗。

13. 谷歌广告：AI Mode 与新广告形态（01:09:21-01:14:37）



场景 13

本节主旨 谷歌广告：AI Mode 与新广告形态

完整内容

曹卿云：下来咱再谈一谈谷歌的广告收入。这怎么看广告业务？一方面，搜索的增速从第一季度的 19% 降到 17%；另一方面，YouTube 的增速从 11% 加速到 13%，感觉有一个分化。这是基数问题，还是大模型重写广告引擎的竞赛里面，谷歌掉队的信号？

周默：跟模型没有太大的关系。我们理解，二季度推荐算法在谷歌广告里面的贡献没有那么强。推荐算法本身也不是一个季度贡献好几个点的，它一个季度能贡献 0.5 个点就非常好，大概率还贡献不到。我觉得谷歌这个季度广告弱，主要还是没有放出来更多的库存。

之前谷歌放了不少库存。这个季度我们看了各种产品的库存情况，也和 agency 沟通，在季度前的报告里就提到，谷歌没有放出来很多库存。相比而言，Meta 有不少库存可以放出来。Meta 的广告这个季度其实还可以，下个季度因为汇率原因稍微低一点；Meta 在 Facebook 的 Reels 里面放出来非常多软广库存，很像 TikTok 和抖音里的软广。但谷歌这个季度没有放出很多库存，因为季度节奏问题，所以会稍微差一点。还有美元指数的影响，美元指数基本上会把这些公司的美元收入拉低 1 个点。

曹卿云：这次财报披露，AI Mode 月活超过 10 亿，但是 AI 界面下单次查询能赚多少钱，谷歌至今没有披露，只给了方向性的安抚：AI Overview 的商业查询变现“令人鼓舞”，单次响应成本降到发布以来最低。AI 查询的单查询收入长期能到传统搜索的水平吗？如果到不了，搜索终局的估值逻辑是不是会重写？

周默：对于 AI 对谷歌广告的影响，现在是比较容易的 low-hanging fruit 已经摘完了。传统广告推三个广告位，但如果算法判断 AI Overview 对广告收入有增量，就可以在后面再加一个 AI Overview 广告位。每扫一次 query，都可以计算加几个广告、怎么调配传统广告和 AI Overview，让这一次 query 的商业化价值最大化。

谷歌广告有非常多长 query，也有很短、没有出现广告名词的 query，这些没有做商业化的 query 应该占到所有 search query 的 60% 以上。但是 AI Max 和 AI Overview 的优势，恰恰是读懂这些长 query，在长 query 里面推出来原来传统广告推不出来的广告。这就是过去几个季度 AI Overview 对谷歌广告带来的贡献。

再往后就要推 AI Mode 对谷歌广告的贡献。AI Mode 的广告不是和传统广告并存，而是一个原生、全新的广告形式。推 AI Mode 广告一定会影响传统广告：如果 AI Overview 曝光变大，影响传统广告的份额，就会影响传统广告的比重。但现在 AI Mode 很多还是增量份额，所以影响还不小。

谷歌这个季度已经有一些好的苗头，开始推一种新的广告形式，叫 **Conversational Discovery Ads**，也就是对话式广告。你和 AI Mode 对话时，它会识别意图，像一个漏斗：购物意愿弱时不推购物卡片；明显感觉购物意愿变强时，就加速推购物卡片和电商广告在对话里的比重。接近最后决策、识别到你就是在买、在选商品时，它会给出很完整的广告商品介绍和广告对比。这是更智能的广告推荐形式，特别适合线索广告和电商广告。

这种形式和现在的广告完全不一样，所以谷歌要找 agency 和广告主推广，也要让广告主理解怎么做 AI Max 的推广，不只是原来直接买关键词，还包括用户适应。这是全新的广告形式，不能指望两三个季度内一下子产生很大营收。它有可能明年贡献 3—5 个点的营收，但大概率不可能今年就这么快贡

献。

画面说明

选中画面为女主持人提问与谷歌广告数据图表，画面可见搜索、YouTube 增速和 AI Mode/AI Overview 相关文字；随后切到男嘉宾解释对话式广告。

14. 算力供需：CAPEX 与 AI ARR 的匹配 (01:14:38–01:19:57)



场景 14

本节主旨 算力供需：CAPEX 与 AI ARR 的匹配

完整内容

曹卿云：展望明年，现在各家 CSP 的 ROI 都很高，算力也很紧缺，所以大家都在拼命投。但是到明年会有大量算力集中投入使用，你怎么推演供需变化？

周默：每年到年底或者股票跌了以后，关于供需变化的讨论都非常多，只不过每次讨论难度都会变大。供给其实代表 CAPEX，需求其实代表 AI ARR，所以我们讨论的是 CAPEX 和 AI ARR 的问题。

这是从这个季度来看越来越难讨论的问题。原来一季 data center 是 350 亿美元时，30GW 的 data center 就是一万亿美元。如果去年年底、供应链还没有大幅涨价时讨论，2027 年的 30GW 对应一万亿美元 CAPEX；我们又看到今年前两个季度 AI ARR 五到十倍增长，这是非常明显的变化。但到了现

在，这一万亿 CAPEX 变成了 **1.7 万亿美元**。所以，到底多少 ARR 才合理、能够 justify 这 **1.7 万亿美元 CAPEX**？

原来一万亿美元 CAPEX 需要投五年。如果每年 CAPEX 都一样，代表未来五年的成本是一万亿美元。但 ARR 增速很快，我觉得至少可以给整个行业 **40%—50%** 的 CAPEX 增长空间。如果按五年算，合理情况是：假设 CAPEX 以后增长不是很快，三年后的 ARR 能跑通今年的 CAPEX，是比较理想的数值。

如果按三年后的 ARR 倒推，未来两年每年增长 **50%**，或者每年 **60%**，未来两年可能更高，达到百分之大几十。那么比如年底的 ARR 能达到明年 CAPEX 的二分之一到三分之一，是一个考虑投入周期、比较容易计算 ROI 的值。一万亿美元 CAPEX 就需要 **3000—5000 亿美元 ARR**，这个数字比较容易实现。

我觉得现在讨论明年年底这几家 labs，加上 CSP、开源模型的 ARR 能不能到 **3000 亿美元**，说服力很强。但如果 CAPEX 到了 **1.7 万亿美元**，就需要 **6000—8000 亿美元 AI ARR**。如果只算现在闭源模型的收入，很难算到这个数。所以供应链大幅涨价推动 CAPEX 变化，最后就造成 ROI 函数里要求的回报越来越高。

现在看 demand 的合理性，不能只看闭源模型营收，还要考虑开源模型营收和各种具体场景。比如简单假设 OpenAI、Anthropic 和几家闭源模型到明年底加起来可能有 **2000—3000 亿美元 ARR**。还有新场景，明年比较大的场景是量化投资，量化的 KPI 很高，也包括开源模型收入。

难点是很难 quantify 开源模型收入。它分散在 New Cloud、比如前两期访谈嘉宾 Fireworks 的收入、企业自己部署的本地化场景里。市场上没人能把开源模型收入合并起来，和闭源模型做 benchmark。全部加起来，明年可能能算到 **6000 亿美元**，但可能需要几个月、更多人统计和更明确的数据集，才能理解算上开源和各种场景后到底能不能达到 **6000 亿美元营收**。因为门槛变高，追踪门槛也变高，算供需的难度就更高，中间的 concern 和底线也会多很多。

画面说明

选中画面为 CAPEX、30GW data center、AI ARR 与三年回收周期的数字推演表，列出 1 万亿/**1.7 万亿美元**投入和 **3000—8000 亿美元 ARR** 区间。

15. 模型与云：Bargaining Power 的再分配（01:19:58–01:24:12）

09 Anthropic 毛利 50% → 85-90%

毛利 才是 真定价权 (营收 ≠ bargaining)		
24 年末	25 Q1	25 Q2
50%	60%	85-90%
待起	对外披露	bargaining 抬升

销售结构：~70% 经 CSP，~90% 经 AWS Bedrock，AWS 抽 40%
意义：毛利上升 → 估值护身

硅谷坐标 周默退场 深度概念图



SILICON VALLEY
VECTOR
硅谷坐标

你已经能显著的看到Anthropic的毛利率增长非常快了

美股回调
资本开支VS ROI之争

微软财报

亚马逊财报

Meta财报

谷歌财报

明年
资本开支

云厂
模型格局

英伟达
财报
展望

场景 15

本节主旨 模型与云：Bargaining Power 的再分配

完整内容

曹卿云：你上次判断模型正在取代云成为入口，感觉 bargaining power 更偏向模型。这个季度微软反复强调模型可替换、Harness 跟模型分离，AWS 和 GCP 也在做多模型平台。如果编排层、上下文、记忆和企业数据都掌握在云厂商手里，bargaining power 是不是又能回到云和应用这一层？

周默：如果没有这么大的开源变化，长期来看 bargaining power 肯定还是在模型公司手里，因为模型公司占主要收入、主要毛利。过去一个季度已经能显著看到 Anthropic 的毛利率增长很快。毛利率本质上反映一家企业的 bargaining power；收入不一定能反映，但毛利率可以。毛利率的价值就是你比上下游多赚钱。

比如二季度，Anthropic 可能从 50%—60% 到现在 85%—90%，已经体现出它在行业里面很强的 bargaining power。同时它的 revenue 从原来的 10 亿变成现在的 20—30 亿，所以这个环节上下游都能赚钱：模型公司、CSP、半导体公司都能赚钱。当然赚最多的是模型公司、存储和英伟达，这三个环节在一个 GW data center 里面 share 提高最多。

但往后看，CSP 有更多开源模型，可以把 router 分配给更多开源模型，整个 share 会有大变化。因为本质上，开源模型在美国的定价权掌握在 CSP 手里，而不是开源模型公司手里；CSP 是开源模型收入的主要 driver。所以 CSP 会有更好的 ROI 和毛利率，单 GW data center 的 CAPEX ROI 比原来和模型公司合作更好。

对模型公司来说 margin 肯定会受影响：竞争力更强的竞争对手出现，就不得不降低 gross margin。对于半导体公司不好说是好是坏。整个环节按毛利率 budget share 看，确实会从模型公司流入开源公司、流入 CSP。

难点是本学期出现的新变化：开源模型公司和 Frontier Lab 之间的差距重新缩短到 3—4 个月。如果一直保持甚至缩短，刚才说的趋势会继续；如果差距重新拉大到 6—9 个月，可能又出现逆向变化。关键就是开源模型和闭源模型的差距保持 3—4 个月，还是 6—9 个月，里面有很多变量。

从 Frontier Labs 角度，参数变化比一年前快。去年底 Frontier Lab 模型差不多 1 万亿参数，到五六月可能已经到 3 万亿参数，往后还可能到 8—9 万亿参数。也就是说不到一年参数规模增加十倍，而且开源模型每个月都在迭代。如果开源模型和闭源模型只差三四个月，CSP 的 bargaining power 就会越来越强。

画面说明

选中画面为模型公司、CSP、半导体/存储之间的毛利率与份额分配表，并显示 3—4 个月和 6—9 个月的开源/闭源模型差距，以及参数规模递增。

16. 英伟达财报前瞻：算力投入与供应链（01:24:13–01:26:01）



场景 16

本节主旨 英伟达财报前瞻：算力投入与供应链

完整内容

曹卿云：最后展望一下英伟达财报。现在各家谈的 CAPEX 加起来已经是 7000 亿美元以上，市场会怎么看？

周默：我觉得更重要的不是单看英伟达，而是 Anthropic 的 IPO、几家 lab 的 ARR。英伟达相对来说估值并不贵，但**存储**、光学等供应链价格上涨，可能会影响市场对利润的判断。

英伟达在 GPU 上的 markup 大概是一比三、一比四。未来会不会像 TPU 一样，把 GPU 和**存储**、网络、数据中心其他部分拆开，重新看每一部分的成本和议价能力？这可能是投资人接下来要关注的事情。现在的 CAPEX 很大，GPU 仍然是最大的单项投入，但供应链价格变化会让整个数据中心的回报率重新计算。

如果 CSP 把更多算力交给开源模型，或者自己直接卖 TPU、GPU，就可能提高自身的议价能力和 ROI；但模型公司要维持领先，又必须继续投入前沿训练。行业面临的是一个可持续性和利润之间的取舍：投入越大，领先模型的能力可能越强，但要让这些投入有合理回报，就需要看到对应的 ARR、利用率和长期需求。

曹卿云：好，感谢周默老师今天把微软、Meta、亚马逊、谷歌以及英伟达财报前的关键问题都拆开来讲。

周默：谢谢大家。

画面说明

选中画面回到女主持人与男嘉宾双人/主持镜头，配合英伟达、CAPEX、**存储**和光学供应链的收束字幕，随后进入节目结束画面。