

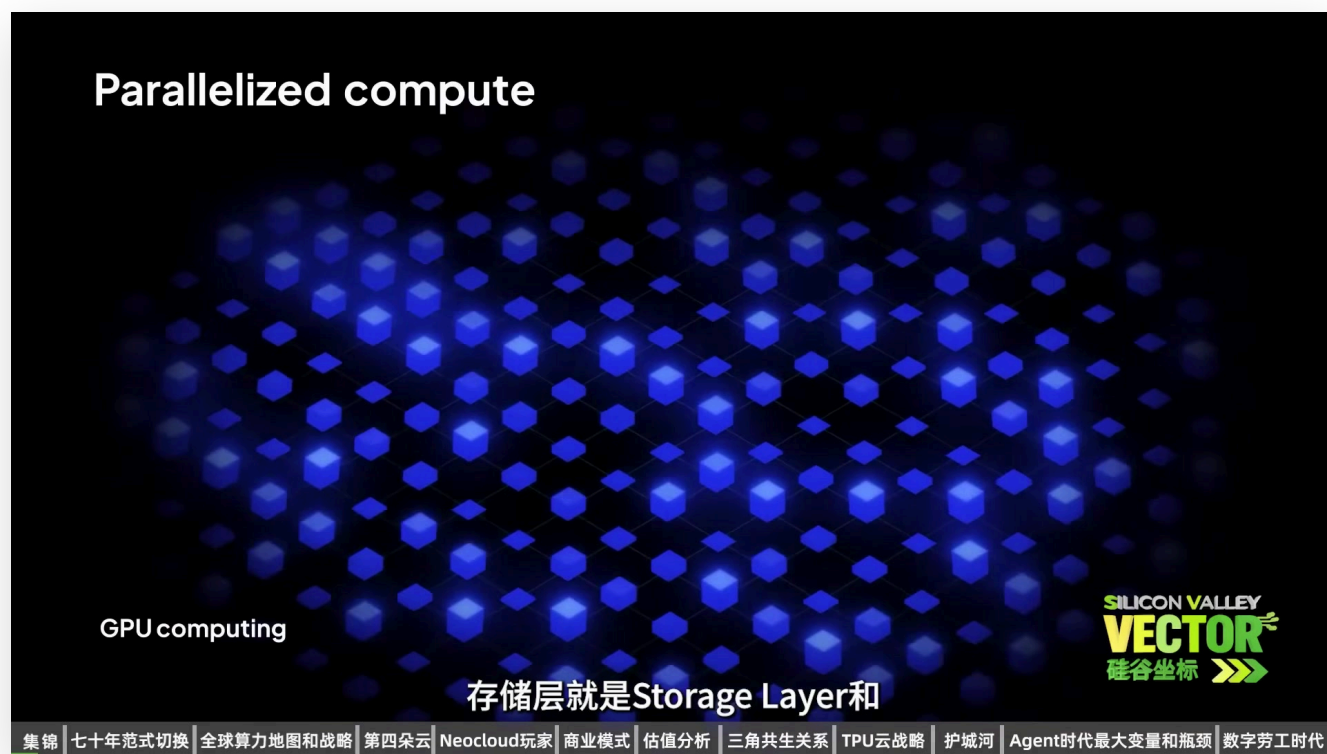
硅谷坐标 x 璞林资本Kenny Zhang:Neocloud崛起背后的供需博弈与AI基建重构 – 图文解说

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 **Neocloud** 展开，Kenny Zhang 将 AI 算力基础设施拆成计算、**存储**、数据库、网络和企业工作负载等层次，解释 CoreWeave、Google TPU、Anthropic 与 xAI 等事件背后的**供需关系**。对谈重点讨论 AI 从训练转向**推理**和 **Agent** 后，**资本开支**、**合同积压**、**GPU/ASIC**、**Storage Layer** 与 **Database Layer** 如何重塑产业链。核心问题不是 Neocloud 是否增长，而是它能否在高资本开支、低利用率和周期波动中建立可持续的利润结构。

01. 算力行业的三件事 (00:01-01:29)



场景 1

本节主旨 算力行业的三件事

完整内容

曹卿云：这个月，AI 算力行业发生了三件事。Anthropic 和 xAI 签了一笔 400 亿美元的算力大单，一家 AI 公司付钱给自己的直接竞争对手买算力。CoreWeave 的合同积压突破 990 亿美元。谷歌联手黑石，要成立一家独立的 TPU 算力公司。

这三件事的背后是同一个故事：一个叫 Neocloud 的新云产业，正在重塑整个 AI 基础设施的格局。它在 AI 产业链中间，做着云巨头不擅长、而 AI lab 不愿意做的事。去年大家讲的是它的崛起，而今年，当 AI 从训练转向**推理**和 **Agent** 时代，未来 Neocloud 能成为 AI 时代的第四朵云吗？还是会向十九世纪铁路基建的方向演绎？

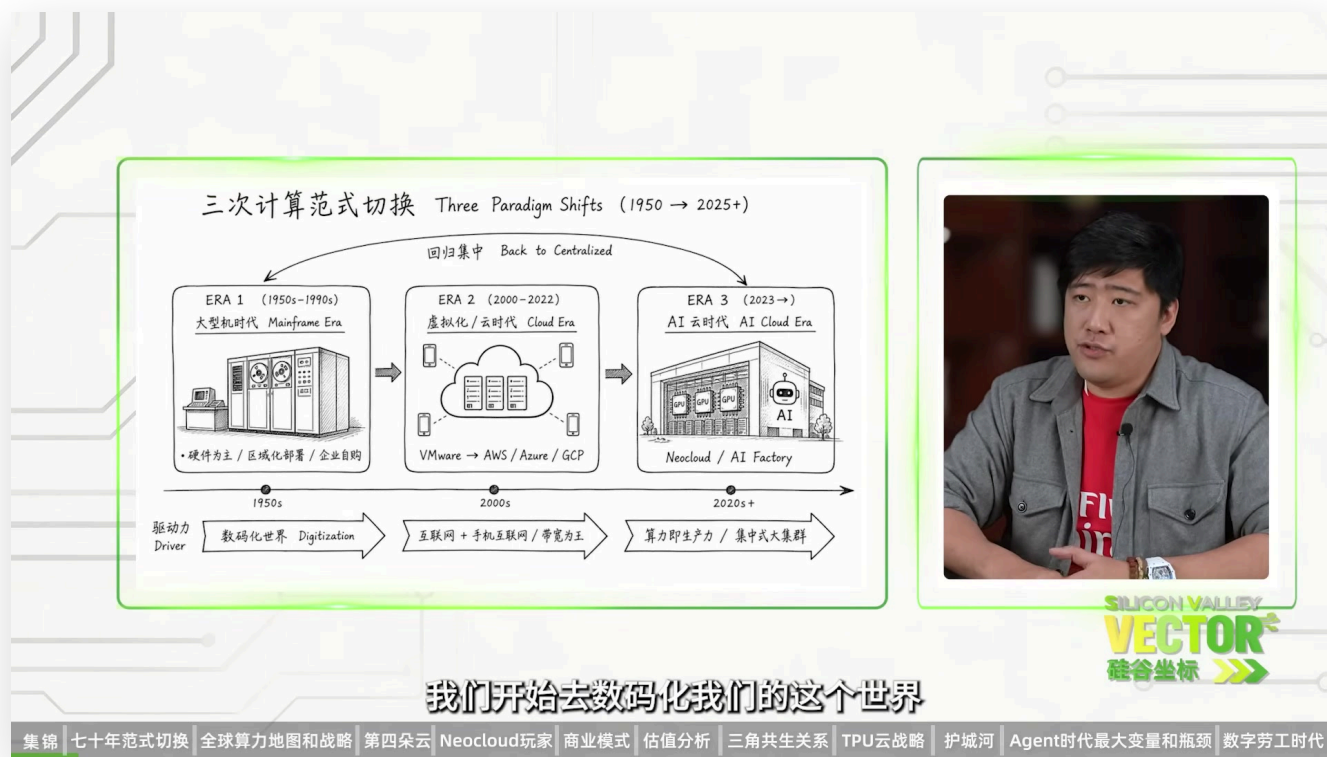
Kenny Zhang：我认为将来一个很大的发展方向是，**存储层**，也就是 **Storage Layer**，和数据层，也就是 **Database Layer**，这两个我觉得会结合在一起。因为机器它不需要去看数据中心，然后再看到这个 Storage 层，甚至它可以直接跟硬件进行交互。就是**存储层**是未来很大的一个层，可以贯穿整个不同的训练也好、**推理**也好、agentic worker 也好、企业的 enterprise worker 也好、To C 的 worker 也好，包括可能是本地部署、端侧。

曹卿云：其实都很重要。今天的嘉宾 Kenny Zhang，Valence 璞林资本首席投资官，会带我们把这张图讲清楚。

画面说明

一张画面显示“Parallelized compute”和“**GPU computing**”，蓝色立方体组成网格；另一张是数据中心机柜通道，字幕写着“数据层就是 **Database Layer**”。画面底部可见节目章节导航和“硅谷坐标 Silicon Valley Vector”标识。

02. 三次计算范式切换 (01:43-06:45)



场景 2

本节主旨 三次计算范式切换

完整内容

曹卿云: Kenny, 你好! 欢迎你来到《硅谷坐标》。

Kenny Zhang: 谢谢曹老师, 欢迎您的邀请。

曹卿云: 今天我们主要是想讲 NeoCloud。然后在讲这个 NeoCloud 之前, 我想先退到一个历史最远处去看这个问题。我们人类从 1950 年第一台大型机开始, 到今天的 Neocloud 管理计算资源, 大概经历了几个时代? 而每一次计算基础设施范式的切换, 它背后又是什么样的力量在推动?

Kenny Zhang: 对, 我们经历过主要是三个比较大的时代吧。第一个呢, 就是从 50 年代到 90 年代末期, 从大型计算机再到民用和商用小型计算机。这个时候主要是以硬件为主, 一般都是区域化或者局部化去部署, 就是企业自己去采购和管理自己的服务器以及小型计算机。第一个大时代, 计算机的

算力是远超于人力的。

最开始的这个时代，就是从八十年代、九十年代开始，我们开始去数码化这个世界，把很多数据以及要处理的文档、文件传送到电脑上，让我们更方便地去运用和使用它们。这个时候出现了第一代的**存储层**、**数据库层**以及**计算层**。

然后第二个时代呢，就是从 **2000 年**以后，一直到 AI 时代的开始，算 **2022 年、2023 年**吧，演变最快的第二个时代，就是虚拟化这个时代。它加速的开始，主要是因为互联网企业的崛起，尤其是到了手机互联网时代。**所以如果您看最开始的这些大云商，最成功的海外就是 AWS，国内的话就是阿里云。**

它之所以成功，是因为它把自己的内部需求以及算力外部化。什么概念呢？海外有 Black Friday，中国有双十一。也就是说，这两大电商，亚马逊和阿里，自己需要匹配非常强大的算力系统，才能在这些需求波峰的时候，给所有用户提供一个比较好的购物体验。**但是平常的时候，它又不需要这么多算力，那怎么办呢？**它就把这个算力虚拟化，卖给、租用给其他企业。这样逐渐形成了最开始的云的雏形。

举个例子，**2010 年**中期，一个 CPU 大概它的算力是 300 TFLOPs 左右吧。但实际上，大部分企业级应用可能只需要 10~20 TFLOPs 的算力。所以在云时代，主要就是说把这些硬件的算力虚拟化，然后再分布式地提供给企业或者用户的 B 端和 C 端需求。

所以在这个云时代，核心竞争力和核心资源诉求其实是带宽。比如头条、抖音的这些视频，一个视频可能要分发给上亿人去看。所以那个时候，这些云主要的核心是去拿带宽的量级，然后再分发给各个客户，B 端或者 C 端的需求。**那个时候对算力的要求并不是特别高，而是说谁能在同一时间整合非常高频的数据调用和分发。**

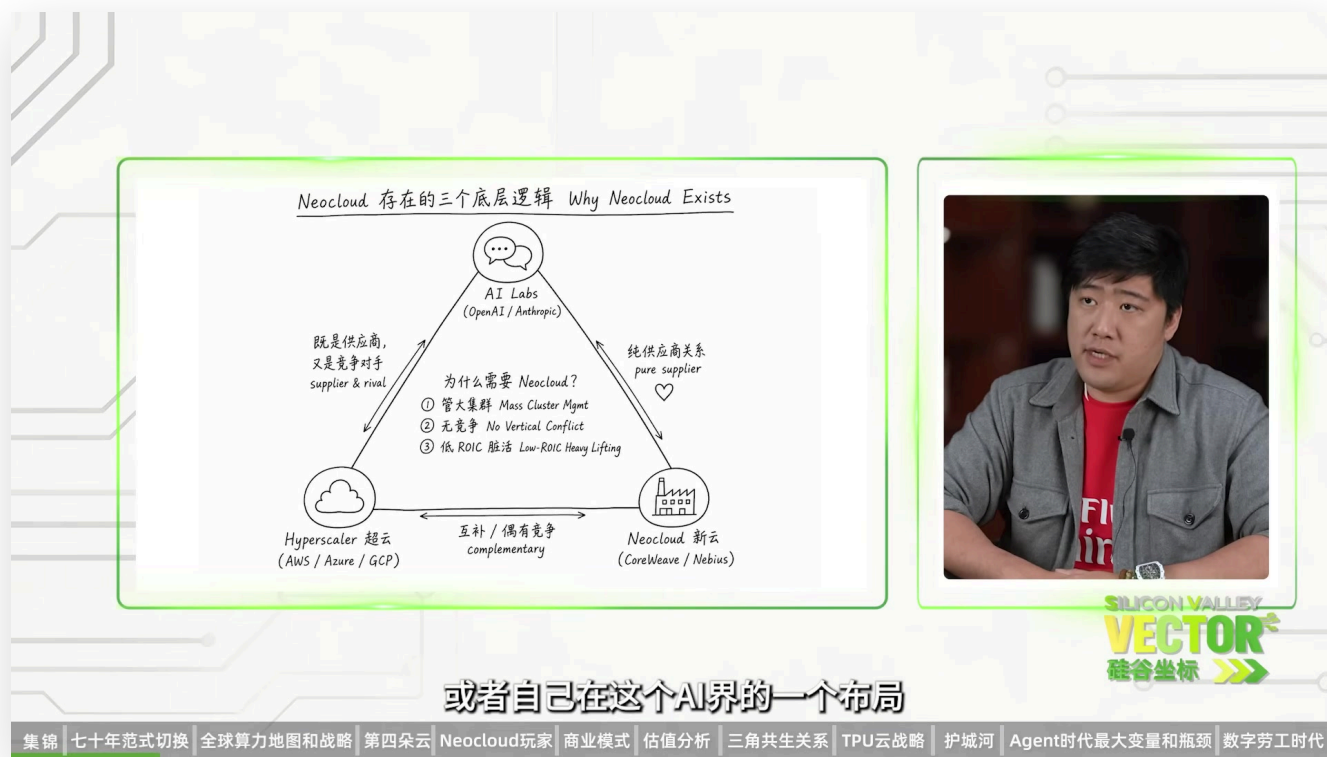
然后到了 AI 时代，就出现了 AI 云，其中比较有代表性的就是 **Neocloud** 里边的 CoreWeave 和 Nebius 这两家公司。我们现在到了 AI 云时代，整个是反过来的，又回到了像五十年代、六十年代大型计算机的时代。什么概念呢？AI 云需要做的不是这种高频的分发，而是需要把算力集中去管理和优化非常大的算力集群。

在 AI 时代，算力的需求可以说是近乎无限的。但它核心的需求不是带宽，而是怎么样去把这些集群持续在一个非常高性能的运算当中。**所以 Neocloud 更像是大型机时代的这样一个范式。**

画面说明

一张图表标题为“三次计算范式切换（1950→2025+）”，并列显示 ERA 1 大型机时代、ERA 2 虚拟化/云时代、ERA 3 AI 云时代；右侧同时出现 Kenny Zhang 的访谈画面。另一个关键帧是 Kenny Zhang 的单人访谈镜头，底部有节目章节导航和中文字幕。

03. Neocloud 为什么存在 (06:45-12:10)



场景 3

本节主旨 Neocloud 为什么存在

完整内容

曹卿云: Neocloud, 你觉得它本身为什么会需要这个新的物种? 它为什么会存在? 它解决了什么传统解决不了的问题?

Kenny Zhang: 传统云和 Neocloud, 它们本身的源头和需要做的事情, 其实基本上可以说是同源不同宗, 就是它的差异性是非常大的。传统云是跟着互联网的需求起来的, 最开始是这样; 然后 AI 云呢, 是从 AI 起步的。

为什么大家发现 AI 的算力需求这么庞大? 那是因为在 AI 时代, 它有几个并发式的算力需求, 就是训练、推理, 还有包括从去年年底开始的 agents 需求, 再加上 B 端企业的推理需求, 它其实在同一时间爆发的。

传统云其实更是一个线性的发展的过程。一开始是在 2C 端，电商、短视频、长视频这块的分发，然后逐渐演变成帮助企业把多余的算力提供给 B 端企业，让它们进行分布式部署。**这样的话，企业端能降成本，不需要自己去管理和采购这些硬件。**

所以在 AI 云时代，我觉得 Neocloud 存在的意义核心有三个大的逻辑。第一个就是我刚才说的，它需要一个非常强的管理和优化超大型密集算力的能力。

第二，因为传统云是从互联网巨头演变而来的，这些互联网巨头又各自在 AI 模型和算力领域有着很多自己的布局。**所以对于新时代的这些 AI 模型来讲，这些超大云商既是供应商，也是竞争对手。**那所以对他们来说，他们希望有一个自己的垂直体系，给他们提供所需的算力。

这一点我觉得比较明显。像美国几家大的云商，微软、谷歌、亚马逊，它们其实都有自己的模型，或者自己在 AI 界的布局，也想在入口端有一席之地。**所以它们对这些新进的巨头，像 OpenAI 和 Anthropic，一直会有拉扯的关系。**

怎么讲呢？OpenAI 最开始可以说是微软扶持起来的。但是微软自己也有内部的算力需求，所以对 OpenAI 来讲，它一直觉得自己在某一个时间点，微软不可能把所有资源都倾向于它，肯定还是有自己的考量的。**那所以 OpenAI 就主要去扶持了像 CoreWeave 这种 Neocloud。**这样的话，它能定制化地得到非常稳定的算力，也可以放心地把自己的 API 以及训练全部放到 CoreWeave 这种 Neocloud 里边。

因为它们没有任何竞争关系，Neocloud 在上层没有应用端的布局，也将来也不会有应用端的布局。

第三个考量呢，Neocloud 的存在性跟企业的资源分配以及能力分配有关。这个其实传统云也是这么去做的。数据中心虽然是一个很庞大的行业，我十五年前就开始看数据中心这个行业。当时在美国的回报率，每年差不多杠杆前是 8%~10% 左右，最初期可能稍微高一点，在 10%~12%，但基本上就稳定在 8% 到 10%。

对于传统云这些互联网企业来讲，它们正常的互联网、软件类或者 PaaS 类业务，ROI 是很高的，可以达到 45%、50%，甚至更高。对它来说，自有资金应该多投放在这些高 ROIC、高回报的应用和主业当中。像这种低回报、又需要特别专业行业经验的领域，它觉得自己没有必要把所有资源都倾向于这边。

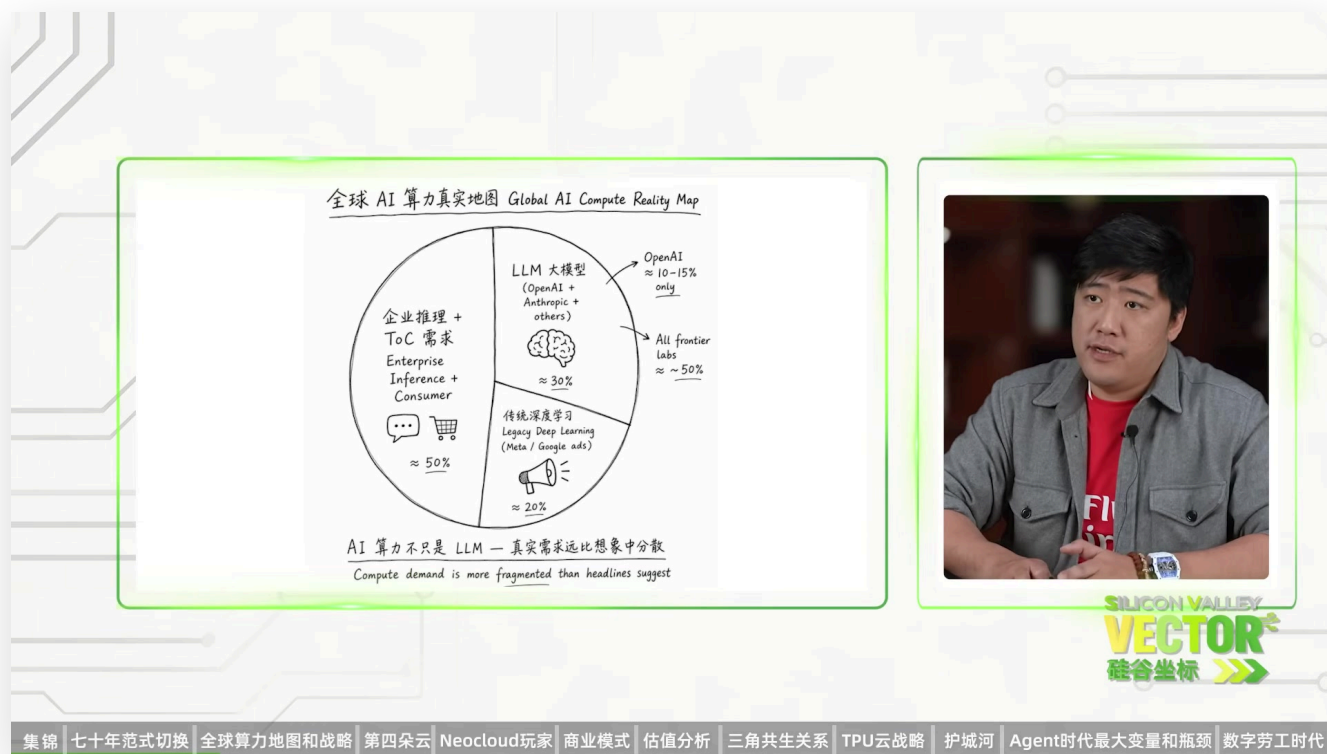
所以在云时代，整体来讲，超大型云商大概占全球、中国以外的数据中心客户的 30% 吧。其实大部分数据中心提供的电力以及制冷供给，还是第三方提供得比较多，大概占 60%~70%。

曹卿云：对，你刚刚讲到，Neocloud 整个行业是因为 AI 算力的需求端驱动的。那今天能帮我们算一下，全球 AI 算力的真实需求到底有多大？

画面说明

关键图表标题为“**Neocloud** 存在的三个底层逻辑”，三角形连接 AI Labs（OpenAI/Anthropic）、Hyperscaler（AWS/Azure/GCP）和 **Neocloud**（CoreWeave/Nebius），中间列出“管大集群、无竞争、低 ROI 重资产”三点。另一张画面是 Kenny Zhang 访谈镜头，字幕提到传统云。

04. 全球 AI 算力的真实分布 (12:11-14:02)



场景 4

本节主旨 全球 AI 算力的真实分布

完整内容

曹卿云：我看到有个报告说，OpenAI 消耗的算力其实只占全球 AI 算力总量的 10%~15%。所有前沿的 AI lab 加在一起，算力总需求可能也只是大概一半的体量。那我们到底怎么样去看整个需求端？有多少不同的需求，他们是谁在用，未来的增速是怎么样？

Kenny Zhang：我觉得这个是很正常的。整体来讲，我们看到前沿 AI lab，加上这些超大云商自己本身的 AI 模型，对算力的需求要分两块，因为 AI 的范围特别广泛。

我们现在泛指的 AI，可能就是基于 Transformer 产生的这些 AI 模型。您刚才提到的 OpenAI 和 Anthropic，还是主要基于大语言模型，也就是 LM 下边的算力。

其实从云时代开始到现在，像 Meta、谷歌这些主要以广告收入为主的互联网企业，一直在用基于 Transformer 做出来的深度学习，也就是 Deep Learning 模型。所以这一块其实占算力的很大一部分，可能有 20% 多。

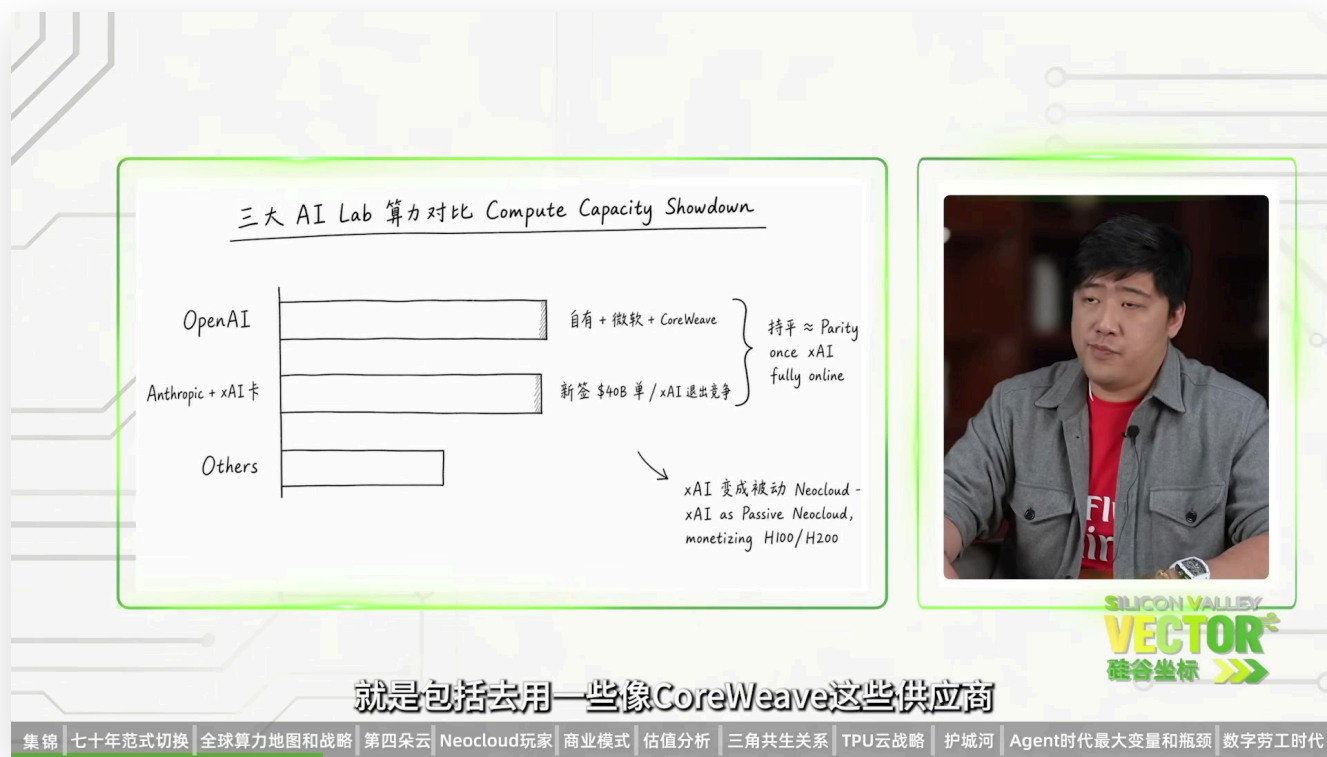
然后 Anthropic、OpenAI 以及其他 LM 模型，现在可能占到了整个算力的 30%。剩下的就是企业推理，以及用户自己的 To C 端需求。所以您刚才提到这些数字，我觉得是合理的，将来大概率也会维持在这样的一个状态。

曹卿云：我们讲到 AI lab 在算力采购上面的这些策略。有的公司我们看到，它可能是更依赖 Neocloud 和这个云巨头 Hyperscaler。

画面说明

图表标题为“全球 AI 算力真实地图”，圆形分区标注 LLM 大模型约 30%、传统深度学习约 20%、企业推理与 To C 需求约 50%，旁边写有“OpenAI $\approx 10-15\%$ only”和“所有前沿 AI lab $\approx 50\%$ ”。右侧为 Kenny Zhang 访谈画面。

05. AI Lab 的算力采购选择 (14:02-18:40)



场景 5

本节主旨 AI Lab 的算力采购选择

完整内容

曹卿云：它的外部产能，比如 Anthropic，有的公司是从主要外购，逐步开始增加自建比例。比如刚刚讲到的 OpenAI，它享有更多自己的控制权。有的可能采用核心自建加弹性外购的策略。你可以帮我分析一下，这些前沿 lab 在战略选择的时候有哪些考量？

Kenny Zhang：今年您看到的情况，其实在公司布局战略上，以及一部分历史原因上，都有一些差异。那我们先拿 OpenAI 来讲。OpenAI 核心的创始人 Sam Altman，从第一天就觉得算力的需求非常庞大，尤其推理这一块，大概率是训练的十倍，甚至更高。那它从 2022 年、2023 年就大量地采购和布局算力。

Anthropic 呢，去年我还记得创始人 Dario 在一个访谈上诟病 OpenAI，说 OpenAI 为什么要无限制地采购这么多算力和硬件。然后今年我们也知道，Anthropic 从去年的大概年底、年化收入 90 亿美元的量级，到了现在可能在 500 亿美元的量级，年底可能会到 1000 亿美元以上。那明显这个需求是远超他们预期的。

所以他们在这两个月明显看到自己的算力是不够的。退一步而言，OpenAI 的策略是对的，所以明显发现 OpenAI 今年在算力供给这一块是非常从容的。Anthropic 现在没有办法，只能在外边大量采购。它都不能说是弹性采购了，现在属于刚性采购，因为算力不足已经影响到它年化收入的增长。

它在年初的时候超过了 OpenAI，甚至接近 OpenAI 两倍的量级。以 4 月底这个时间点来看，从这个时间点开始，我认为 OpenAI 有机会去追赶上 Anthropic，因为它们在算力方面的不足是很有前瞻性的。

所以我觉得未来的格局，大概率就像您的问题里提到的：它们部分自建，部分去第三方采购。核心考量是，OpenAI 不见得是最有效、最能优化这些大集群的，所以它们一定要在外部，包括使用 CoreWeave 这些供应商，这样能对比和考量自己内部对算力以及 token 产生的优化程度。

最后一部分就是弹性采购。按照新的需求爆发，它们一定要实时地把算力跟上，不然收入以及用户体验会受到影响。

曹卿云：这几个头部 AI lab 现在拥有的算力差距有多大？

Kenny Zhang：Anthropic 最近签了一笔非常大的订单，跟 xAI 签的，因为 xAI 基本上算是退出了 AI 大模型竞争的角色，但他们手里边还是有大量的 H100 和 H200 卡，也就是算力。xAI 等于是被动式地把自己变成了一个 Neocloud，但他们大概率也不会继续发展这个业务，等于把自己手里的 GPU 和算力变现，然后提供给 Anthropic。

目前我们看到的情况是，当这些 xAI 的算力全部上线以后，OpenAI 和 Anthropic 的整个算力供给差不多是持平的。

曹卿云：未来的空间，你觉得还会增加吗？对于 Neocloud 来说？我的理解是，AI lab 如果自己自建的话，它们整个的增长天花板会受影响。但未来的趋势你觉得，对它们来说空间还是会很大？

Kenny Zhang：是，我觉得空间很大。但是这些 Neocloud 需要把自己的战略以及商业模式想清楚。现在 Neocloud 我觉得分成两个派系。

画面说明

关键帧显示“三大 AI Lab 算力对比”条形图：OpenAI 标注自有、微软与 CoreWeave, Anthropic 加 xAI 卡标注新签 **400 亿美元**订单，旁边写有 xAI 退出竞争后成为被动 **Neocloud**、变现 H100/H200；右侧为 Kenny Zhang 访谈画面。

06. Neocloud 的两种商业模式 (18:41-27:11)

Neocloud 三大派系
看似同一类, 实质完全不同

第四朵云	CoreWeave · Nebius · FluidStack 长合同锁定 + PaaS 增值 毛利 25% (硬件) + 15-20% (PaaS) → 整体利润率 30% 出头 Kenny: “第四朵和第五朵金花”
裸金属派	IREN · Iris Energy · Crusoe 多为 Crypto 挖矿转型而来 纯硬件租赁, 按时计费, 受价格波动 + 资金双重挤压 Kenny: 持中性态度
混合派	Oracle 同时做 AI 大单 + 抢超云份额 双线作战, 资源丰厚 战略未明, 待观察

那他们基本上在我的视角里

SILICON VALLEY
VECTOR
硅谷坐标

集锦 | 七十年范式切换 | 全球算力地图和战略 | 第四朵云 | Neocloud 玩家 | 商业模式 | 估值分析 | 三角共生关系 | TPU 云战略 | 护城河 | Agent 时代最大变量和瓶颈 | 数字劳工时代

场景 6

本节主旨 Neocloud 的两种商业模式

完整内容

Kenny Zhang: 比较有代表性的就是 CoreWeave 和 Nebius。这两个很明显是想去跟传统云抢未来整个大云环境中的位置，他们想做第四朵云和第五朵云。

什么概念呢？传统云，比如拿 AWS 为例，它的利润率大概在 40% 上下。那这 40% 的利润率为为什么这么高？如果只看硬件来讲，利润率应该只在 25%~30% 之间。它能把利润率提升 10~15 个点，是因为它们提供了很多 PaaS 层面的服务，包括存储、分发、加带宽、数据库，以及云这一边的很多管理。软的这一层占收入的 30%~35%，是非常高毛利的业务，毛利可能达到 80% 甚至 90% 以上，所以把整体利润率提升了很多。

现在来看 **Neocloud**，它们也分成几派。CoreWeave、Nebius 也是想在硬件基础上提供很多附加值，就是刚才提到的 PaaS 层面。但 AI 和传统云提供的服务不一样，附加的 PaaS 也很不一样。

目前我们看到 **Neocloud** 在发力的，是整个算力的优化、KV cache 的合理部署和调用，再加上它们会做一些最近 OpenAI 和 Anthropic 都在布局的 Serverless RL，也就是强化学习训练。

如果您看 CoreWeave，今年到年底，**存储**、KV cache 这一块应该给它带来的年化收入差不多到 **3 亿美元**了。我觉得将来会到 10 亿、**20 亿美元**。如果这些 Neocloud 能把 PaaS 层面的东西做到收入的 **15%~20%**，我觉得肯定做不到 **30%~35%**，因为 AI 大模型公司和其他大型超云公司自己有很强的 PaaS 能力，但做到 **15%~20%** 就已经很不错了。

这样的话，它们整体利润率应该能达到 **30%** 出头。因为 AI **GPU** 纯硬件层面的利润率大概在 **25%** 左右。

另外还有很多 Neocloud，大部分像 Oracle、SoftBank Energy，还有美国一些稍微小的，以及从当年挖矿、crypto 转型的 Neocloud。比较有代表性的就是 IREN、Iris Energy，还有现在一些私有公司，Crusoe、FluidStack。它们现在主要做裸机柜、裸金属租用。**Oracle 可能想两边都发力，在真正的超云市场做角逐。**

但这些 crypto 转型的公司，在我的视角里基本就是裸金属，更像硬件租赁业务。我觉得这个商业模式很容易受到价格回报以及资金供给层面的挤压。**所以这一类商业模式，我持比较中性的态度。**

曹卿云：刚刚我们在讲需求端，未来 AI 往**推理**端和 **Agentic** 时代继续发展，那你觉得未来对于整个现在云的格局会有些什么变化？

Kenny Zhang：我觉得大概率 CoreWeave 还有 Nebius 应该能成为第四朵和第五朵云，因为它们的布局 and 理念比较超前。**当然它们两个也各自有不太一样的商业模式，但核心是我们从传统云时代进入到以硬件为核心基础的 Neocloud 时代。**

未来我觉得这两个都会回归到一条线上，也就是说 AI 原生云。大家要为所有新的 agentic workflow 去做流程上的整体改变。现在很多不同的企业，包括英伟达、CoreWeave、超云 Hyperscaler 本身，加上一些之前比较小、特别个性化的云商，像 DigitalOcean，都提过 AI 原生云，也就是 AI native cloud。但大家对这个词的理解并不太一样。

有一些像 DigitalOcean，更偏向以产品理念来看 AI 原生云，最近发布了 15 款还是 20 款给 AI 原生云做准备的产品。然后英伟达和 CoreWeave 这两个绑定比较深的战略伙伴，它们所说的 AI 原生云，就是以 AI 算力大集群的管理为核心。它们讲的是一个比较粗的概念，这个是 AI 原生云。

像超云这些公司，基本不太提这个概念，因为它们觉得之前做的事情和现在做的没有本质区别，只不过 B 端 workflow 会从人来使用软件、数据库、存储层，转到机器来代替很多人使用软件。它们觉得没有特别大的差异化，所以不太提那么多 AI 原生云。它们的理念是继续做一直做的事情，只不过面对的对象不同了：以前主要面对人，现在主要面对机器。

曹卿云：那你觉得未来 Agent 时代的需求，更多会被这些 Hyperscaler 抢走市场份额，还是会给这些新的玩家提供更多机会？

Kenny Zhang：市场份额，我觉得还是在计算层，或者说物理层。最后的市场份额，还是要看它们客户、下游能拿到多少市场份额。简单来说，哪一个模型将来的市场份额更大，它的上游供应商市场份额就会随着它走。

它的市场份额分两块：一块是需求端，就是跟它绑定比较深的客户，未来发展的这些模型公司，也包括超云；另外就是它自己的资源能力。这两年发生了很多数据中心，因为这些云商的上游是数据中心，数据中心把我们叫作 PowerShell，只提供电力和制冷。它需要上游企业正常、稳定地提供最基础的电力链设施。

还有一个就是卡这一块，这是它们核心的资源能力：能不能拿到最先进以及最合适的卡。所以这是资源端目前决定 Neocloud 市场份额的因素。它是供给端和需求端都并存的。

刚刚我们聊了很多需求端的问题，现在来看看供给端。供给端行业现在有几个比较重要的玩家。按照 2025 年收入口径，一个是 CoreWeave，51 亿美元的收入，可以占到整个公有云市场 22%。

画面说明

一张“Neocloud 三大派系”图表将 CoreWeave、Nebius、FluidStack 放在“第四朵”一栏，说明长合同锁定与 PaaS 增值；IREN、Iris Energy、Crusoe 在“裸金属派”；Oracle 在“混合派”。另一张是 Kenny Zhang 访谈画面，字幕提到从人使用软件转向机器使用软件。

07. 供给侧玩家与 TPU 增量 (27:11-30:24)



场景 7

本节主旨 供给侧玩家与 TPU 增量

完整内容

曹卿云：之后 Nebius 大概只有 5 亿美元的收入，Lambda Labs 也是大概 5 亿。这样合计的话，可能长尾加在一起的需求还会有 100 多亿美元。能不能给我们讲一下现在整个 Neocloud 的供给侧，这些玩家各自有什么差别？

Kenny Zhang：您刚才给到的数字基本上是去年的。其实今年，或者未来两三年的格局，变化还是很大的。尤其是 TPU，从 2025 年大概只有一百多万，可能 100 万、120 万、130 万颗的供给，到了今年 400 万颗。明年按照我的预测，可能会上升到 1000 万到 1200 万颗，量级已经接近英伟达了。

亚马逊的 Trainium 芯片，明年也会到 400 万到 500 万颗的量级，今年可能 200 万颗，明年甚至可能更高。所以这几家基本上都是英伟达系的，它们 100% 的芯片以及集群部署都是以英伟达为基础的。我觉得 GPU 的崛起，带动了 FluidStack 这种 Neocloud。

所以我觉得未来的格局，三大巨头应该是 CoreWeave、Nebius，然后是 FluidStack；最后亚马逊自己也会以训练为主，在这边占有一席之地。但因为亚马逊是超云，我们先把它放开不谈。

Lambda 我简单讲一下，它更侧重于编程。我觉得这是未来另外一个方向，就是做垂直化、区域化的 Neocloud，专业性特别强。我们现在主要聊的还是大规模、范式化、泛应用、泛算力的 Neocloud，FluidStack 其实比较像，所以我就把这两个合在一起说。

当我提到 Nebius 的时候，您可以把它想象成 FluidStack 就是一个用 TPU 版的 Nebius。所以 CoreWeave 和 Nebius/FluidStack 的本质区别是这样的：CoreWeave 选择了最传统的基建模式，也就是说它去拓展资源，拓展资源之后，跟下游客户，Anthropic 也好，OpenAI 也好，或者包括一些新进的算力需求大户，像 Jane Street 这种高频交易公司，还有 HRT，在这一块 CoreWeave 选择去签长单。

因为一个 GPU 的商用生命周期大概是 5~6 年，它们就把这 5~6 年全部锁给一个客户。就是说如果用了之后还有残值，那属于锦上添花。所以 CoreWeave 的模式就是去拿到这些芯片的资源，去组成自己的大集群。

画面说明

两张关键帧都是访谈镜头：一张为曹卿云，另一张为 Kenny Zhang；画面可见节目水印、章节导航和中文字幕，没有额外图表或姓名条。

08. CoreWeave 与 Nebius 的融资路径 (30:25–35:34)

为什么 Nebius 选择低杠杆 Why Nebius Avoids Leverage

创始团队基因 = Yandex 出身
/ Internet-Native DNA

银行杠杆 Bank Debt
Almost None

客户预付 Customer Prepay
40-50%

合同打折 20-25%
Contract Discount

杠杆前 IRR 35-40%
Pre-Leverage IRR

微软 + Meta 大单
MSFT + Meta Deals

选择承担价格波动风险
而非财务杠杆风险
Takes price volatility risk,
not financial risk

SILICON VALLEY
VECTOR
硅谷坐标

集锦 | 七十年范式切换 | 全球算力地图和战略 | 第四朵云 | Neocloud玩家 | 商业模式 | 估值分析 | 三角共生关系 | TPU云战略 | 护城河 | Agent时代最大变量和瓶颈 | 数字劳工时代

场景 8

本节主旨 CoreWeave 与 Nebius 的融资路径

完整内容

Kenny Zhang：然后我把这个大集群的整体算力打包给一个头部客户。既然我有了很大的合同价值，大数的话我们算，它每花一块钱，大概在未来 **6 年**能赚回来 3 块钱。它采取的是用锁定价格，然后用财务上的杠杆来提升回报率。

这是非常传统的高速公路、数据中心，包括传统商业地产、商业基建的模式。它的路线非常清晰。然后它自己的融资成本，随着银行以及投资人对这个行业的接受度越来越高，下降得非常快。

它在 **2023 年**发行的第一笔贷款，利率是 **14.5%**，算是 high yield 这样的债务发行，非常高的成本。它今年发行的新一轮贷款，利率只有 **6.2%**，而且已经成为投资级债务，也就是 investment grade。

所以它的 IRR 提升了相当多。杠杆前 IRR 大概是 **23%~25%** 左右，三年前杠杆后提升不大，可能在 **35%~40%** 左右，甚至更低一点，在 **30%~35%**。现在它的杠杆后回报率已经稳稳超过 **50%**，非常可观。

对比传统基建项目，基本上杠杆前回报是在 **5%~6%** 左右。我们说的核心铁路、码头、高速公路，杠杆后可能达到 **10%~12%**，好一点的会做到 **15%**。我大学毕业以后，老东家就是伯克希尔，是现在管理 **2 万亿美元** 规模、全世界最大的商业地产和基建公司。他们是一个有一百年历史的加拿大资产管理公司，一直采用这样的模式，非常稳定的现金流，也是非常高可预测性的回报。

然后我们看 Nebius 和 FluidStack。Nebius 选择另外一条路线，也就是说它更愿意承担价格方面的一定波动。Nebius 的团队是之前俄罗斯最大的搜索网站 Yandex 的创始团队，他们把公司转移到阿姆斯特丹，也就是荷兰，完全把自己脱离了俄罗斯的整个系统。他们 also 把 Yandex 卖掉了，然后用 Yandex 卖掉的资金，开始做 Nebius 的新发展方向，也就是 **Neocloud**。

他们非常有互联网思维，对传统互联网和传统云也有很深的理解。所以在他们的视角下，他们希望承担一定的市场算力价格波动，但不愿意承担财务杠杆的风险。

所以他们最近几个大单，微软和 Meta 这些单子，在财报法说会上提到，预付金达到了 **40%**，甚至 **50%**。他们自己只发行一些可转债，也就是在公司层面发行可转债，整体杠杆率不高。

当然，如果想让客户预支大量 Capex，也就是**资本开支**，也要牺牲一些合同价值。我们测算下来，Nebius 在合同价格上应该打了大概 **20%~25%** 的折扣。但它的 ROI 也非常可观，因为没有杠杆，所以我们只聊杠杆前的收益率。因为有非常大额的前期预付款，它们的杠杆前收益应该能达到 **35%~40%**。

曹卿云：所以从创始团队的基因来看，Nebius 选择不用杠杆的形式，但同时也说明它有更高的上杠杆空间，是吗？

Kenny Zhang：但我觉得它们应该不会上很多杠杆，应该一直用 **40%~50%** 的预付款形式继续发展。我之前所说的所有回报率，都是看它的净现金流的。

曹卿云：嗯，因为 EBITDA 里边是不包括折旧的嘛。刚刚你讲的商业模式，我可不可以简单理解成，这个商业模式是借钱去买 GPU？

画面说明

图表标题为“为什么 Nebius 选择低杠杆”，显示 Yandex 创始团队、几乎没有银行杠杆、客户预付 40%~50%，以及合同折扣 20%~25%、杠杆前 IRR 35%~40%。右侧为 Kenny Zhang 访谈画面。

09. 折旧、合同与盈利曲线（35:34–45:02）



场景 9

本节主旨 折旧、合同与盈利曲线

完整内容

曹卿云：然后是把 GPU 的时间租出去，再用租金还这个债。有一个比较大的疑问：GPU 本身是在贬值的，英伟达的新产品一到两年就会出来新一代，产品也会有更替，使用寿命其实也是短短几年。感觉像是用一个会贬值的资产去借一个长期的债，这样的结构逻辑上是不是能够站住脚？

Kenny Zhang：我觉得这是现在市场层面以及很多人对于这个商业模式的一个误区。就是说，如果咱们先看传统基建，码头、铁路，包括数据中心，其实它们也有很大的折旧，只不过折旧期限可能更长一些。数据中心的设备折旧，这些电力系统，大概是 15 年到 20 年左右，所以每年大概有 5%~6% 的折旧。

如果假设 AI 芯片的使用年限是 **5 年到 6 年**，其实这跟传统云没有那么大的区别。就是说在 CPU 时代，当然最近这几年折旧年限一直在逐渐延伸，最开始传统云出来的时候，它们 CPU 的折旧年限也是 **5~6 年**，到现在可能有一些超云商把它拓展到了 **7 年甚至 8 年**，跟 AI 云差别并不大。

无论是基建、传统云还是 AI 云，它们借到的负债，其实对应的不是 GPU 硬件本身的残值，而是整体合同的价值。所以对于 CoreWeave 和 Nebius 来讲，因为合同主要都已经签了 **5 年到 6 年**，基本上把芯片的商用价值用到了合理范围。如果在这个基础以上还有一些使用空间，对它们来说就是锦上添花，基本上都已经折旧完了，能多用一天就多赚一天，刨去电以后是 **100% 的利润**。

所以在传统云、新云、基建里，它们算的账是整个合同，刨掉所有电力运营、一定的研发和销售成本，最后按照采购成本去算回报率。银行也是算这个大账。这也是为什么过去三年 CoreWeave 的利息成本能下降这么多，主要是大家现在熟悉了这个商业模式，觉得它跟传统云没有太大差别。

现在对这个商业模式有很大质疑的，其实是二级市场。因为从财务报表上来看，CoreWeave 是比较有代表性的一个公司：去年 Q2、Q3，它的利润率还在 **12%~14% 的水平**；去年 Q4 下降到 **8%**；今年上半年断崖式下跌到 **1%~2%**。这增强了市场对这个商业模式的质疑：怎么越做利润率反而越下降了？尤其还是在 GPU 现货市场里，第一季度现货价格还上升了 **40%**，那为什么它的利润率下降了？

其实这个很简单。我十几年前开始看数据中心的时候，也出现过同样的情况。在需求爆炸式增长的情况下，尤其又锁定了未来五六年的现金流，它肯定要大规模建设。因为需求爆发相对突然，就像 **2015 年、2016 年中国手机互联网和云计算爆发的时候**。比较有代表性的，就是我之前作为首席战略官的公司，叫万国数据。那个时候万国数据 **2016 年**上市，刚开始报表也非常不好看。

但核心原因跟 CoreWeave 现在一样，Nebius 也一样，就是它们现在在建的规模大于收租的体量。我们把它换算成数字，可能更直观一些。CoreWeave 在上一个季报公布，现在已经通电的部分是 **1 个 GW**，整体布局的电力是 **3.5 个 GW**，在建大概率有 **1 个 GW**。

在它已经通电的那 **1 个 GW** 里，我们测算下来，可能只有 **80%**，也就是 **800 个 MW**，是真正产生收入的。因为通电、客户上架和它自己把服务器部署起来，还是会有两三个月的差异。所以它这 **1 个 GW** 里只有 **800 个 MW** 在赚钱，有 **200 个 MW** 通了电但没有赚钱。然后另外它还在部署和建设，今年大概有 **1 个 GW** 的规模。

也就是说，它在建的和收租的，超过收租规模的一倍。这样就会把利润率带下来。如果展望到今年年底，它大概率会有 1.8~2 个 GW 在收租，还是会继续建差不多 1 个 GW。这样的话，从在建超过已有收入的一倍多，到年底在建和已经收租的只占 **50%**。

到年底，收租这一部分的利润率应该在 **25%** 左右；它现在如果再建，到收租达到 **50%** 体量的年底，利润率应该在 **12%~14%** 的水平。到了明年年底，大概率已上架、在收租的是 3 个 GW，还会继续再建 1 个 GW，这个时候在建只占现有规模的 **30%**。利润率就会更接近稳定，明年会到 **20%** 出头。

然后再叠加我之前提到的 KV cache、RL 这些高利润的 PaaS 服务，利润率在 **2028 年底**很有可能上升到 **25%**。在建和收租之间的差距缩短，最后可能奔着 **30%** 去。到那个时候，资本市场应该会提前意识到，新云其实跟旧云差别不是特别大了，无论从芯片折旧层面还是盈利模式，实际上都非常接近，估值应该也会接近。

现在的估值差非常大。传统云是在 **20~25 倍**利润率，增长大概在 **30%** 左右；新云按照我们的测算，只有 **10 倍**稳定利润，但利润增长超过 **100%**，甚至未来三年的复合增长率是 **120%~130%**。

为什么新云的估值倍数会比传统云低？我觉得长期来讲，新云应该比传统云稍微低一些，因为它**资本开支**的量非常大。按照每 MW 来算，CoreWeave 和 Nebius 的**资本开支**大概在 4500 万到 **5000 万美元**，传统云每个 MW 大概是 1500 万到 **2000 万美元**，甚至更低。如果用自己研发的芯片以及白牌网络、带宽设备，可能下降到接近每 MW 1000 万到 **1200 万美元**。

第一，资本投入力度要比传统云大。第二，利润率大概永远达不到传统云 **40%** 的水平。因为传统云把很多 PaaS 打包业务卖给企业，可以收取高昂利润，但未来也许会受到 AI 模型和应用的冲击，类似现在有些应用层软件受到自己做 web coding，或者找小型软件开发商做出原来大型软件公司打包出售的部分功能的冲击。

将来我觉得它们的利润率会逐渐 converge。传统云可能下降到 **30%~35%**，Neocloud 上升到 **25%~30%**，相对接近但还是会差一些。整体估值应该会有一些折扣，但不应该像现在这么大的折扣。

曹卿云：对。你刚刚讲的 CoreWeave，它现在的收入可能就只有 **50 亿美元**左右。

画面说明

一张图表比较传统云与 Neocloud：传统云 P/E 20~25 倍、利润增速约 30%、每 MW Capex 1500~2000 万美元，Neocloud 约 10 倍、利润增速 120%~130%、每 MW Capex 4500~5000 万美元；下方标注长期利润率区间。另一张是曹卿云访谈画面。

10. 积压订单与效率提升 (45:02–55:08)

为什么 CoreWeave 选择高杠杆
Why CoreWeave Embraces Leverage

利率
Interest Rate:
14.5%
↓
6.2%
Investment Grade
投资级

签 5-6 年长单
Lock 5-6 Year Deals

组大集群
Build Mega Cluster

拿芯片
Get GPU

\$1 投入
↓
6 年回 \$3
杠杆后 IRR > 50%

传统基建打法 — 高速公路 / 数据中心 / 商业地产 — 100 年被验证
Classic infra playbook validated for 100 years

SILICON VALLEY
VECTOR
硅谷坐标

是没有任何关联性的

集锦 | 七十年范式切换 | 全球算力地图和战略 | 第四朵云 | Neocloud玩家 | 商业模式 | 估值分析 | 三角共生关系 | TPU云战略 | 护城河 | Agent时代最大变量和瓶颈 | 数字劳工时代

场景 10

本节主旨 积压订单与效率提升

完整内容

Kenny Zhang：然后它今年的收入指引是 120~130 亿美元左右，但是 Capex 要花出去的钱大概在两三百亿美元。它的积压订单现在是 900 多亿美元，大家对这个积压订单也会有一些不一样的看法。

曹卿云：比如 CoreWeave 最大的单一客户就是微软，占到它 2025 年收入的大概三分之二。其实还有大量来自 OpenAI。你对这种积压订单的质量，以及它背后带来的风险怎么看？

Kenny Zhang：我觉得这是资本市场、二级市场对于 CoreWeave 最大的一个误区。您可以把它想象成，它现在这个时间点的收入，跟这个时间点的资本开支没有任何关联性，因为它是瀑布型结构：今天花一块钱，未来六年会收回来三块钱。

所以今年看到的 120 亿、**130 亿美元收入**，是它之前两年投放和上架的结果；今年花的 300 亿甚至 **350 亿美元**，是未来六年会收回来 900 亿到 **1000 亿美元的利润**。要把这两块拆开看。最简单的看法就是，它现在花一块钱，未来 **6 年**赚 3 块钱。

如果假设**资本开支**不大幅提升，它的收入和资本开支会逐渐交叉，最后应该在 **2029 年**收入超过资本开支，然后利润率直接到现金流这一块。所以这是一个非常核心、市场还没有用来看的方式，因为它不是 AI 模型公司。AI 模型公司的资本开支和收入之间时间差比较短，今年花多少，未来 6~12 个月可能就开始兑现或者转成收入，因为它们需要尽快转成收入。

但对于 CoreWeave 和 Nebius 这些企业，只要客户不存在大幅度的财政问题，它基本上就是把这三块钱的利润拆成 5~**6 年**。所以只要建好、服务不间断、客户不出现现金流问题，它的现金流就会一直累积叠加。

我觉得所有基建行业一定要向前看。它的积压订单，是未来需要交付的算力。去年 **2025 年**，它接近 **70% 的收入**都来自微软，作为中间商提供给 OpenAI。但这个收入反映的是它在 **2023 年、2024 年**投放的结果。如果向前看，它的客户背景会非常多元化。

以现在这个时间点看，它最大的客户会变成 Meta，因为 Meta 占它积压订单的 **35%**；微软应该占 **20%**；OpenAI 差不多也 **20%**；Anthropic 可能在 **8%~10%**；Jane Street 现在也到了 **8%~10%**。所以从现在这个时间点，只要这些客户不取消订单，它几乎也不能取消订单，这些订单还是具有一定财务绑定性的。再加上只要能正常、稳定地交付积压订单的需求，三到四年后收入就非常多元化了。

还有一个关键点经常被投资界忽略：它跟英伟达的绑定关系。英伟达承诺，它很大一部分芯片如果采购之后不能部署，英伟达可以把它回收过来。**所以它的财务风险敞口其实没有大家想象的那么大。**

曹卿云：这里还讲到另外一个需求侧的担忧。DeepSeek 如果让 AI 算法效率持续提升，对 Neocloud 的这些积压订单会有什么影响？你怎么看？

Kenny Zhang：我觉得大家可能太关注现货市场，因为今年现货市场确实涨了很多，但这其实不太符合以前传统云的逻辑。

传统云的商业模式和逻辑是，以前企业客户自己采购和管理服务器以及需求，有了多余算力也没有办法弥补成本，所以大家逐渐往云上搬。说白了，就是零售变成批发，或者说集采。**大型超云商用规模、庞大技术以及自己主业的现金流，部署和建立核心骨干算力系统。**因为规模效应，它们在设备、

带宽、服务器和数据中心部署上逐年优化，算力成本每年下降得很快，然后把一部分成本下降的利润留给自己，一部分给客户。

这样客户会拓展出更多需求，把更多原本自己部署的需求搬到云上。这一直是过去十几年云计算发展的核心底层逻辑。AI 云其实也想效仿这个逻辑。

我之前提到，AI 是多方位需求爆发，训练、**推理**、C 端、B 端、agentic workflow、企业商用 workflow，它是并发的。再加上上游算力链，我们在 AI 时代开始时经历了最惨烈的半导体下行周期。因为在 Covid、新冠期间，上游建设了大量供给，当时 PC、手机需求都是倍级增量，而且大家不确定世界供应链和病毒对全世界影响的持续性，所以进行了大量扩张。

大量扩张的结局是，Covid 比预期更快结束。现在耳熟能详的企业，包括海力士、闪迪、铠侠，在 **2022 年、2023 年**几乎算是濒临破产。最惨烈的半导体下行周期，又迅速接上了可能是这 **20 年到 30 年**最大的上行周期。所以它们这几年不敢大刀阔斧地扩张，怕重蹈五年前濒临破产的覆辙，这个记忆还很新。

今年算力市场的现货价格是不正常的。**GPU** 是很伟大的一个产业，但它所做的优化和算力效率提升，我觉得是正常的。效率提升和 token price 下行，是必要条件，这样大量软件、设备以及端侧需求才能蓬勃发展。

这样 **Neocloud** 和 AI 算力公司也能做同样的事情：把一部分优化效率的提升留给自己，再把一部分留给更下游的应用，以及其他 B 端和 C 端用户，这才是健康的生态。包括上游半导体、网络、服务器设备，也会随着对周期确定性的增加，合理扩张。

如果看明年相对于今年，大家可能预计云商、超云商、AI 算力和 Neocloud 的资本开支整体增长 **50%~60%**，甚至更高。但这 **50%~60%** 里，大部分是元件和芯片，尤其**存储**涨价造成的。整个硅，也就是芯片的量级，其实没有很高的提升。

曹卿云：这个资本开支是不是那么一个健康的情况？

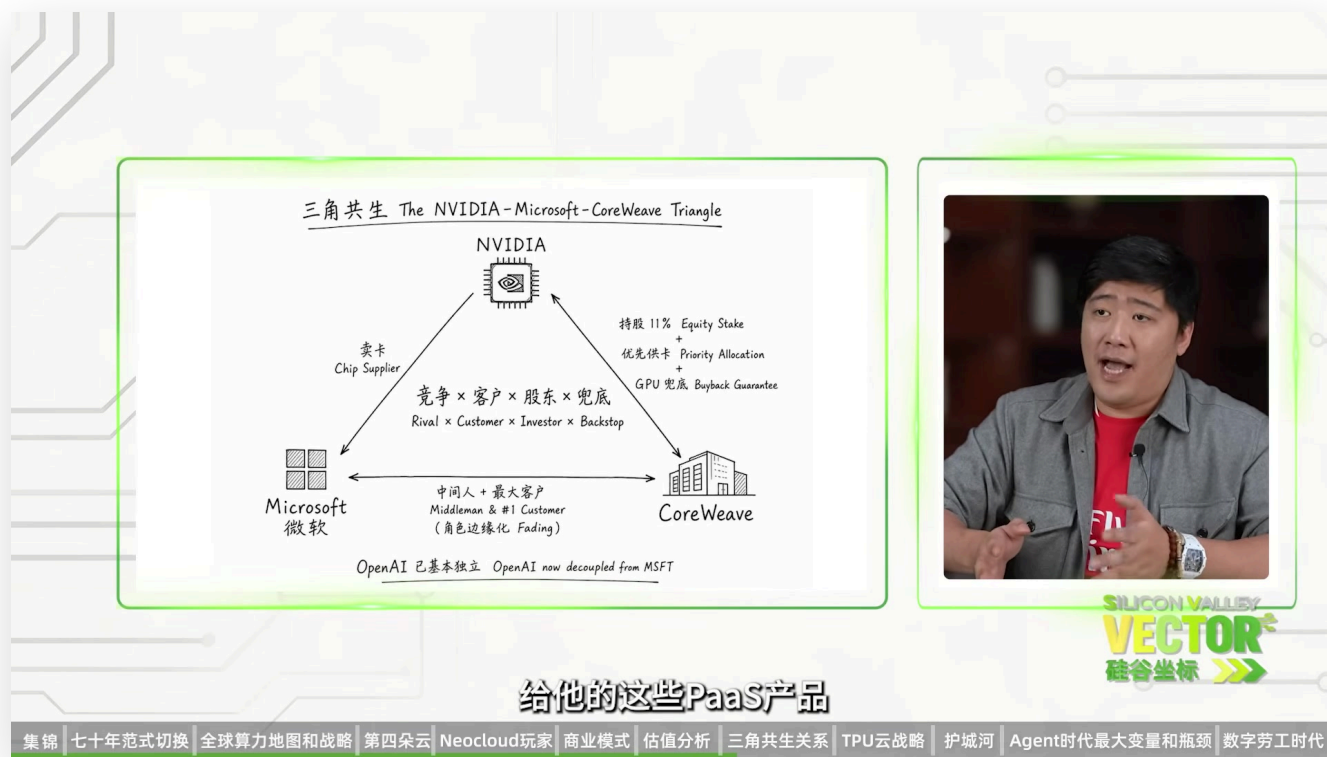
Kenny Zhang：我希望看到的市场，是每年资本开支的增加能跟现金流匹配，再跟 silicon content 匹配。最好每年资本开支增加 **50%**，silicon content 增加 **30%~40%**，可能有一部分是上游加价造成的。再加上下游算力优化、token cost 继续大幅下降，这样真正的需求才会看到真正的爆发。

曹卿云：你讲的就是那个杰文斯悖论，然后成本打下来以后，需求端会有更多的应用爆发。

画面说明

关键帧显示“为什么 CoreWeave 选择高杠杆”图：5~6 年长单、组大集群、14.5% 降到 6.2% 的利率、1 美元投入 6 年回收 3 美元、杠杆后 IRR 超过 50%。另一张是曹卿云访谈画面。

11. NVIDIA、微软与 CoreWeave (00:55:08–01:03:54)



场景 11

本节主旨 NVIDIA、微软与 CoreWeave

完整内容

曹卿云：应用爆发。接下来我想聊一聊整个产业链里面的这些角色：微软、CoreWeave，还有英伟达。微软是全球第二大云服务商，也是 CoreWeave 最直接的竞争对手，但同时它也是 CoreWeave 最大的一个客户。英伟达卖芯片，同时又有 CoreWeave 11% 的股权。刚刚你也提到，如果 CoreWeave 的 GPU 租不出去，英伟达可以给它兜底。那这三者的关系，他们是竞争、客户，还是共生？

Kenny Zhang：这一块可能是我个人的观点，可能稍微重一些。其实我觉得在这个体系当中，微软的角色已经比较边缘化了。

微软在这个体系中的重要角色在一开始。为什么它作为 OpenAI 和 CoreWeave 之间的中间人？**首先两点：第一，微软当年给 OpenAI 做了大量投资，希望 OpenAI 的 ChatGPT 或其他模型，给它的 PaaS 产品，尤其以 Microsoft 办公软件为核心的产品，以及 AI 功能 Copilot，提供极大的赋能。**

对于 CoreWeave 来讲，OpenAI 当时收入和支出的差比很大，有极大的负现金流。**所以微软作为中间人，对整体来讲是一个增信作用。**就是说，如果没有微软，OpenAI 这边能拿到的价格可能极高。

因为基建行业说白了是一个客户企业级信用的逻辑。**比如你的客户是 Triple B 或者 Double A，那你的基建信用等级可能比它差一点，因为你还有自己的运营和布局风险。**但基本上，基建项目以单一项目来看，它的投资信用度，也就是 credit rating，跟客户强相关。

客户信用差，价格就会相对较高，所以微软起到这样的作用。我觉得从近期，一个月前或者两个月前，微软跟 OpenAI 的合作到了一个段落。OpenAI 会继续赋能它的 Copilot，但已经不是一对一的关系了，OpenAI 已经基本上从微软体系独立出来了。

所以我们主要谈英伟达和 CoreWeave 的战略关系。我觉得最核心的是 A、B 两方。从英伟达角度来讲，它想核心扶持和支持自己的一套硬件基建云生态体系。虽然其他超云公司也采购英伟达芯片，但它们肯定不想被一家公司决定未来整个布局和发展。

未来大型超云公司和大型互联网企业，包括头条、Meta、阿里，都会积极自研自己的芯片。英伟达也肯定要在自己的体系内树立护城河，所以想找一家本身能力比较强、又比较互补的公司。**我认为英伟达觉得 CoreWeave，包括 Nebius，在大集群管理和优化方面有独特经验和见解，所以英伟达投资了这两家公司，并且选择性地优先给它们提供芯片和服务。**

还有一层更深的逻辑：传统云商不想跟一个整体解决方案绑定。**如果大家都用标准化 AMD 或 Intel 服务器，那怎么在价格或者服务方面跑出优势？其实比较难。**

传统云从 **2012 年、2013 年**高速发展开始，三家核心超云 Microsoft、AWS、谷歌，把硬件大量白牌化。最早企业用 Cisco 的交换机和 router，到了云时代，硬件大量白牌化。服务器也几乎不像 OEM 那样，从 T0 到 T10 做一个完整整合。OEM，比如戴尔，负责最后 T8 到 T10；台湾原厂商，也就是 ODM，从 T0 直接做到 T8。

这些超云在传统云时代，希望尽量往原厂走，把自己的设计和采购理念放进去。不重要的东西用便宜的，重要的东西，比如带宽和网络设备，就选最好的。因为那个时代带宽、网络和分布式分发很重要。**这样它们能按照团队能力跑出差异化，互相竞争，所以都有很大的采购和研发团队去优化服务**

器。

英伟达一开始，以及其他逻辑芯片厂商，属于做发动机，再配一些核心轴承；ODM 可能把整车做出来，OEM 在上面贴牌、喷漆、改装，再加一部分利润卖给下游客户，大概是这样的生态。

传统云更像喜欢自己改造车的人，只买一个发动机。就像我们小时候看《头文字 D》，买一个 GTR，但自己做很多改进和优化，把车的性能、发动机性能提升到最强，希望比别人跑得快。

到了 AI 时代，英伟达的理念是：不要自己做自主改装，不要自己抖机灵。它说我的整套系统就是最好的、最优解，我就是一直为 AI 算力时代研发的。相当于你应该买我这辆超跑，什么都不用改。

所以它一直想找一家云商，把这个理念完整输出出来。CoreWeave 在一些集群当中就采用这个理念，用整套英伟达网络设备，不做太多硬件层面的自有优化，只做 PaaS 层面的优化。相当于它跟英伟达是合作伙伴，做出英伟达想展现给世界的真正的**推理**和训练能力。

英伟达一直觉得，其他超云厂商最后改来改去，效果不如它理念型的方案。对于 CoreWeave 来讲，有了英伟达的投资和合作关系之后，拿取芯片、服务器应该有一定优先级；英伟达也可能给它在软件、PaaS 层面提供支持。因为英伟达自己也开发了许多 AI 训练和**推理**工具，希望有一个渠道把能力推给更下游客户。

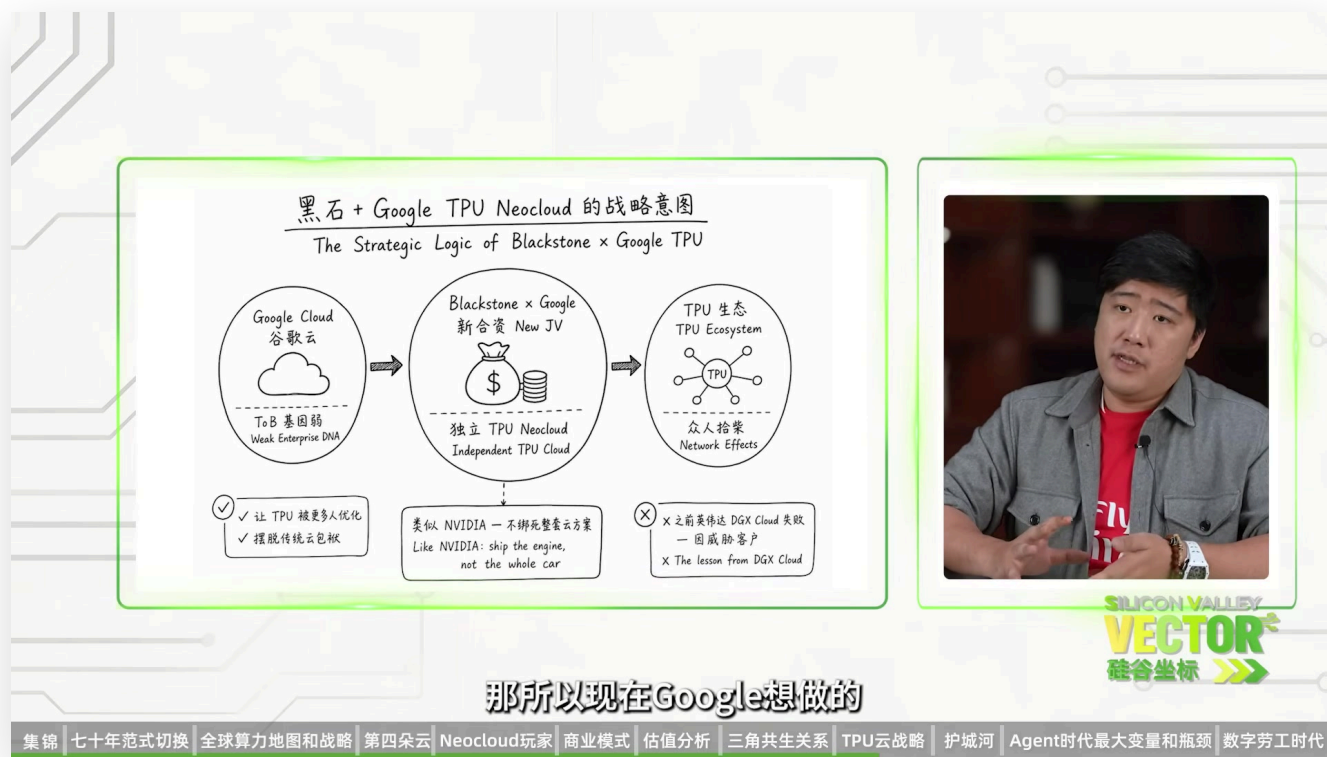
当然资金也非常重要，有英伟达这样的客户背书。包括我之前提到的，英伟达还说，如果芯片没有交付出去，它愿意回收。

曹卿云：还有一个新的新闻，就是黑石和谷歌成立了一个 TPU 的 Neocloud，供给端的谷歌联合了一个金融资本。

画面说明

图表标题为“三角共生 The NVIDIA—Microsoft—CoreWeave Triangle”，三角三点分别是 NVIDIA、Microsoft、CoreWeave，标注卖芯、竞争/客户/股东/兜底、持股 11%、优先供卡和 GPU 回购保证；下方注明 OpenAI 已基本独立。另一张为 Kenny Zhang 访谈镜头。

12. Google TPU 的新云策略 (01:03:54–01:07:37)



场景 12

本节主旨 Google TPU 的新云策略

完整内容

曹卿云：然后用 **Neocloud** 的模式来经营算力。我们知道谷歌已经有自己的 Google Cloud，为什么还要在自己的体系之外再建立一个独立的 TPU **Neocloud**？这对 CoreWeave 有什么影响？因为感觉**推理**时代到来之后，CoreWeave 又在跟英伟达绑定，**推理**芯片以及其他各种各样的推理芯片都出来了。跟英伟达绑定会不会成为它们的劣势？

Kenny Zhang：我跟 CoreWeave 管理层交流比较多，我并不觉得他们认为自己跟某一家是完全独家的状态。但它们内部确实觉得英伟达的芯片是最好的、最优解的。

我们先说黑石和 Google TPU 的联合。如果您的硬件被更多人使用、被更多场景使用，它会有很大的生态效应。这也是为什么英伟达在 TPU 的实用性在这两年被大量提升之后，比如 **2023 年**，角逐的几乎就是英伟达和 AMD，当时还有英特尔，但英特尔的 **GPU** 很快就凋零了。

为什么英伟达生态一直比 AMD 生态好很多？就是因为英伟达芯片被超云厂商和 AI 模型公司大量优化和应用，也包括 DeepSeek。因为它自己的 CUDA 生态，实用性非常强，效率非常高。

所以现在 Google 想做的就是让大家把 TPU 用起来，众人拾柴火焰高，大家都会帮它优化，也可能提升它整个体系。但如果它只输出自己的 Google Cloud 生态，因为一个云继承了自己前世今生，传统云里很多技术和 PaaS 不见得会被新一代 AI 原生公司接受。

所以它不想被自己之前的生态和理念限制芯片层面的发展，而是非常有战略性地选择黑石以及其他资金深厚的财团，把 TPU 推向所有人。

如果反过来看，两年前英伟达也尝试自己做云，它的 DGX 当时也有一点想做云，但没有人愿意用。大家觉得，如果这个云发展快了，将来你可能不卖我芯片了，把芯片留在自己手里，只出售算力，这对大家是威胁更大的路径。所以大家第一天就反对这个路径，没有接受它做云的愿景，英伟达也就不怎么再提这件事了。

谷歌我觉得是一样，只不过稍微反过来：它已经有一个庞大的云生态，但不见得所有人都想依托它云的所有东西。也许很多人想用 TPU，觉得它很好，大家希望选择性地使用你生态中的一部分。

就像宝马、奔驰一样，我可能认可你的发动机，但不认可你的车的设计理念、内饰，我还想自己做。这样谷歌就给大家拓展性，是非常有前瞻性的。

曹卿云：对，还想问问你，怎么看待 Neocloud 真正持久的护城河？有时候我们会想到十九世纪铁路的例子，有很多铁路公司把铁路，还有包括火车，拿去抵押。

画面说明

图表标题为“黑石 + Google TPU **Neocloud** 的战略意图”，从 Google Cloud 经由 Blackstone × Google 新合资指向 TPU 生态，标注独立 TPU **Neocloud**、众人拾柴、摆脱传统云包袱，并对照 NVIDIA DGX Cloud 失败的客户阻力。右侧为 Kenny Zhang 访谈镜头。

13. 重资产行业的护城河 (01:07:37-01:12:50)



场景 13

本节主旨 重资产行业的护城河

完整内容

Kenny Zhang: 最后公司破产了，基础设施留下来了，但是这些公司都没了。

曹卿云: 你觉得未来 Neocloud 的持久护城河是什么？它以后是像你说的变成云，还是也可能随着这个时代没有利润，然后消失，或者成为利润率最低的生意？

Kenny Zhang: 这个类比我觉得比较恰当，而且这个类比其实可以应用到大部分重资产行业。我们之前在伯克希尔的时候开个玩笑：酒店是一个风险特别高的资产类别，尤其在商业地产里边。我们经常开玩笑说，第一个酒店所有者大概率要亏钱或者破产；第二个 owner 一般会打平；到了第三个 owner 才开始赚钱。

铁路也好，基建也好，都是这样。您看当时伯克希尔在 2008 年、2009 年的时候，因为一家澳大利亚最大的基建资产管理公司 Babcock & Brown 破产了，伯克希尔就以资产价值大打折扣的价格，收购了澳大利亚整个铁路、码头、基建类可能 20%~25% 的市场份额，一把全都收过来了。

这个就是因为巴菲特和布朗之前对整个周期的理解，以及对自己在极端情况下承压能力的一个错配，给了 Brookfield 一把占据庞大市场地位的机会。

所以无论铁路，还是后来的住宅、基建，原理其实都是一样的。重资产、基建类、地产类企业破产了，除非烂尾，只要建完了，基本上都会留在这个世界上。因为经济在发展、人口在发展，它在某一个时间点总会填平供给和需求之间的差异。

当供给和需求差异特别大的时候，持有资产的人大概率会有很大的财务损失。所以最好是成本越低越好，供给和需求的差异越小越好。因为它们都用杠杆，基建基本上得用杠杆才能有一个比较合理的收益率，同时它也能加杠杆，因为有稳定的现金流。

但这就像银行放贷一样，它对自己的财务模型和下游需求，一定要有特别保守的预期。当然太保守赚不到钱，规模也起不来；太激进的话，步子跨太大，有可能濒临破产。

所以我觉得 Neocloud 也好，传统云也好，铁路也好，地产也好，最后取决于公司自己的运营能力，以及财务风险敞口。

基建类从来不会一家独大，因为这几乎不可能。大家核心竞争力的差异化没有那么大。说白了，如果您有钱，对收益率预期极其低，就可以在基建行业有一席之地。别人需要十几个点的收益，如果我的收益只想要 2%~3%，或者 4%~5%，又不想加杠杆，那我买了之后可以一直持有这个资产，别人很难撼动我。所以它肯定不会是一家独大，这不可能。

未来的发展，我觉得这个行业肯定有一些会被淘汰，有一些过于激进的，我这里就不点名了。它的产能供给会被更优质的企业接手。就像 Babcock & Brown 这个例子，过于激进企业持有的资产，最后基本都会流向真正财大气粗的企业，或者流向更好的管理者。对社会和资本市场来讲，这是比较良性的发展。

所以我觉得 AI 云也是一样，这些资产会存在。但我们也聊过，芯片不会持续这么长。铁路的折旧可能是 40 年、50 年，AI 芯片 6 年以后残值可能就比较低了，更新迭代可能真的会稍微快一点，所以这个资产不会永久持续在那里。

算力链现在实在是百花齐放，两只手已经数不过来了。未来两三年，规模超过 10 亿美元的 Neocloud 公司，肯定有一大半要被淘汰。第一，你要受到三朵云的影响；第二，可能受到新进新贵的影响，就是 CoreWeave、Nebius、FluidStack；然后还要取决于下游客户的影响。

这个时代也不是所有应用和 AI 模型都会存活到十年以后，有一部分会存活，有一部分不会。如果下游客户不行，也会导致一部分上游供应商，也就是这些 Neocloud，逐步优化和淘汰。但我认为这是健康的，也是必然发生的。

画面说明

画面为曹卿云的单人访谈镜头，右下角有“硅谷坐标 Silicon Valley Vector”标识，底部有节目章节导航；没有可读的图表或额外姓名条。

14. PaaS 层与存储层 (01:12:51-01:16:09)



场景 14

本节主旨 PaaS 层与存储层

完整内容

曹卿云：再想问一个展望未来的问题。不管是从**推理端**，还是 agentic workflow 未来需求上升之后，你觉得整个云基础设施里面，最大的变量会在哪一层？

Kenny Zhang：未来最大的瓶颈会在哪里？我个人认为，最有趣的一层其实是在 PaaS 这一层。

以前在 IaaS，也就是基建这一层，我觉得大家基本平起平坐。哪怕理念上有一些差别，最后结果基本都一样。差别是什么呢？比如微软想做小而轻的数据中心布局，有很多很多节点；亚马逊做大型数据中心部署，然后以非常强的带宽能力和 containerization 能力输出硬件系统。

最后它们两个在硬件层面、物理层面，差异化其实没有那么大。差异化特别大的其实在 PaaS 层面。

微软有先天优势，因为它有很多 SaaS 能力，也有 PaaS 能力，所以可以把很多 SaaS 和 PaaS 放到云上，一起并行销售给你，打一个大包，全都一起推给你。

亚马逊一开始的核心客户可能偏 startup、创业企业为主。创业企业比较省，能省就省，所以会像选菜单一样定向式选择一定功能。亚马逊是把菜单做齐了，但没有想一起都推给你，没有想让你吃 omakase，你自己点。所以它的功能非常丰富，亚马逊在传统云这块做得非常好。

但是无论微软还是亚马逊，为什么我觉得 Google 在传统云时代做得实在一般？因为它没有 To B 的基因。除了广告业务以外，它没有太多企业服务的基因。

所以亚马逊和 Azure 这两巨头的竞争，最大差异化就在 PaaS 层面。但它们提供的都是与人服务的東西，很多数据库、可视性的云管理软件，其实都是给人提供的。

如果我们现在讨论机器或者芯片代替很多知识型劳动力，也就是说很多 PaaS 系统对于机器来讲没有太大意义。举个例子，我认为将来一个很大的发展方向是，**存储层**，也就是 **Storage Layer**，和数据层，也就是 **Database Layer**，这两个我觉得会结合在一起。

因为机器不需要去看数据中心，然后再看到 Storage 层，甚至可以直接跟硬件进行交互。所以我觉得未来核心的层面是 **Storage Layer**。**存储层**是一个未来很大的层，可以贯穿整个不同的训练也好、**推理**也好、agentic worker 也好、企业的 enterprise worker 也好。

画面说明

这一场关键帧是 Kenny Zhang 访谈画面，字幕出现“**存储层就是 Storage Layer**”，另一个下方信息条写着“下一代**存储层**的真正赢家”；画面没有可读的图表。

15. Storage Layer 的竞争 (01:16:09–01:19:25)

The diagram illustrates the Storage Layer as the cross-cutting core. It shows a central box labeled 'Storage Layer 存储层 + Database Layer 数据层 (融合 Merged)'. Above this box, an upward arrow points to 'Database Layer / Attack Upward' with icons for databases and SQL/Vector DB. Below the box, a downward arrow points to 'Optimize Hardware / Attack Downward' with icons for SSD, magnetic tape, and ESSD. To the left of the central box, text asks '为什么? Why?' and explains that answers are cached to avoid recomputation. To the right, it lists use cases: Training, Inference, Agentic Worker, Enterprise Worker, ToC Worker, and On-Device. At the bottom, it states that machines don't need a human UI. A video frame on the right shows Kenny Zhang speaking. The bottom of the slide features a navigation bar with various topics.

下一代存储层 = 跨场景核心层 Storage Layer = The Cross-Cutting Core

↑ 向上吃 Database Layer / Attack Upward
数据库 SQL / Vector DB

为什么? Why?
答案被大量存起来,
不是每次推理都重算
Answers cached,
not recomputed
every time

Storage Layer 存储层
+
Database Layer 数据层
(融合 Merged)

贯穿所有场景 Spans every use case:
✓ 训练 Training
✓ 推理 Inference
✓ Agentic Worker
✓ Enterprise Worker
✓ ToC Worker
✓ 端侧 On-Device

↓ 向下吃硬件 Optimize Hardware / Attack Downward
SSD · 磁带 · ESSD

机器不需要为人优化的 UI / Machines don't need human UI — they talk to hardware directly

他是因为推理的速度变快了

SILICON VALLEY
VECTOR
硅谷坐标

集锦 | 七十年范式切换 | 全球算力地图和战略 | 第四朵云 | Neocloud玩家 | 商业模式 | 估值分析 | 三角共生关系 | TPU云战略 | 护城河 | Agent时代最大变量和瓶颈 | 数字劳工时代

场景 15

本节主旨 Storage Layer 的竞争

完整内容

Kenny Zhang: To C 的 worker 也好, 包括可能是本地部署、端侧, 其实都很重要。

为什么我不谈硬件这一块? 因为硬件供给可多可少, 硬件固然重要, 但将来它的价值还取决于几家之间博弈的情况。

但是软件层面, 我觉得谁能把这一层站住是非常重要的, 就是 PaaS 这一层。其实有一个想法: 是不是在现在或者未来某一天, 我们已经把想问的问题全都问过了? 那个时候, 是不是在 LM 这个层面, 就从纯显性的问答层面, 已经不需要更多**推理**算力了?

您看现在很多数据源已经接近上限，也就是说我们有可能今年或者明年，基本把很多问题都问了几万遍。这个其实也能看到，大模型公司现在回答速度越来越快。我不认为是因为**推理**速度变快了，因为新的卡其实也没有出现。我觉得是因为很多答案已经存起来了。

最后**存储**的模式，下边用什么？硬盘、磁带、ESSD，其实都无所谓。所以这就是为什么我避开硬件不谈，我现在只谈 **Storage Layer**。

我觉得这非常核心。谁能高效、高频、大量、低价地优化**存储**信息和答案这一块，我觉得这其实是一个非常大的机会。

一些传统企业在传统 **Storage Layer** 这一块可能做得不太好，包括很老的企业，戴尔里面的 EMC、NetApp、Pure Storage，这些都是上市公司或上市公司的一部分。还有一些未上市的，我觉得大家可以关注，比如 Vast Data，它一直沉淀了很多 low latency storage 的技术。

所以我觉得还是非常值得关注。如果问我现在能看得到的、在我认知范围内比较有意思的方向，Storage 还是非常 interesting 的，因为它上可以攻击 **Database Layer**，下可以优化硬件类。最后硬件的解决，我们走一步是一步，可能就是纯成本的考量了。

以前高速存储需求很低，因为企业的存储需求就是把一个文件存进去，可能这辈子也不再看一次，所以不在乎高速提取和存储。在 AI 时代之前，Vast Data 主要是高频交易公司使用它，把它作为客户。

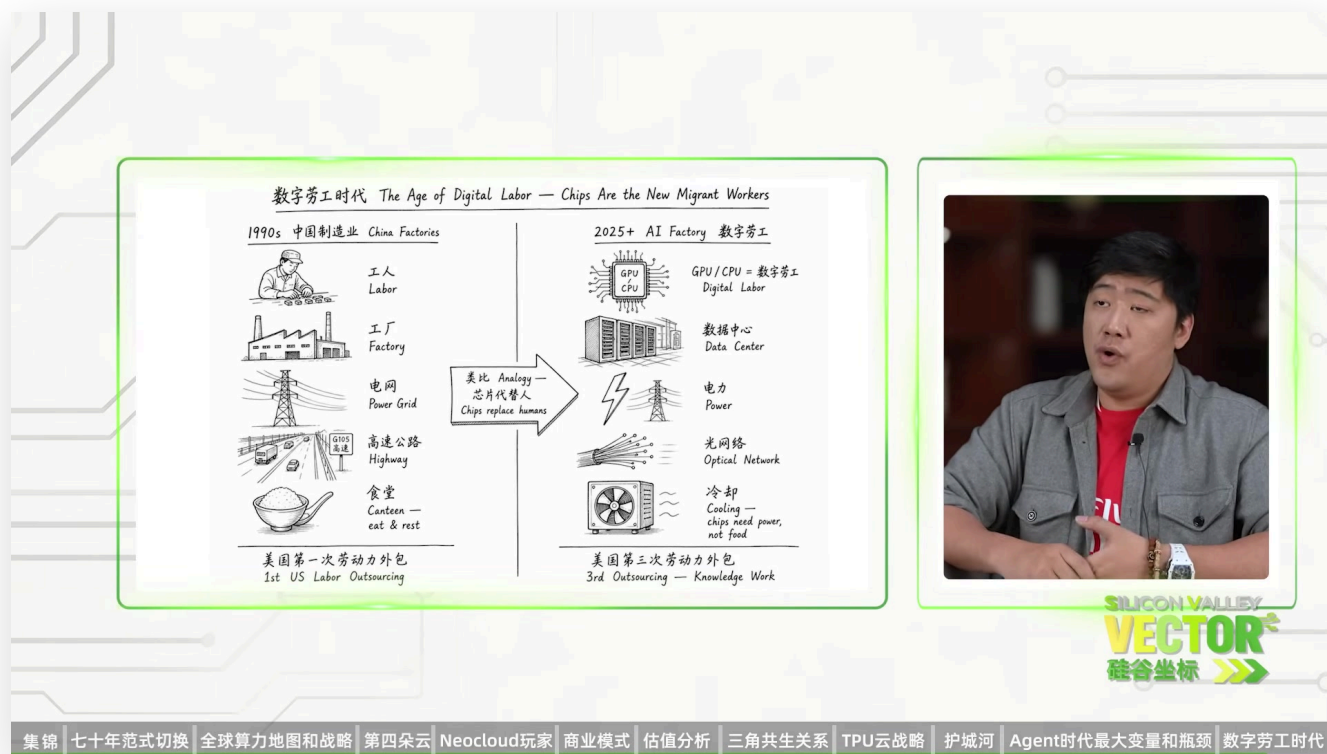
但现在基本上这些头部的，无论超云、AI 算力公司还是 **Neocloud**，都是它的 backlog，它的 backlog 很强很强。所以最后谁能把 Storage Layer 做出来，还有待观望。但我觉得这肯定是一个我们非常关注的赛道。

曹卿云：今天最后一个问题，也想聊一聊你最近公开发表的一个言论：你把 AI agent 比作数字劳工，把 **Neocloud** 比作维持这些数字人的运转的基础设施。

画面说明

图表标题为“下一代**存储**层 = 跨场景核心层”，中心把 **Storage Layer** 与 **Database Layer** 合并，向上连接数据库/SQL/Vector DB，向下连接 SSD、磁带、ESSD，右侧列出训练、**推理**、**Agentic Worker**、Enterprise Worker、To C Worker 和端侧；另一张为 **Kenny Zhang** 访谈画面。

16. 数字劳工与 AI 工厂 (01:19:26-01:27:26)



场景 16

本节主旨 数字劳工与 AI 工厂

完整内容

曹卿云：运转的基础设施。你说，如果芯片在取代人，我们怎么样让数字人感到开心？你说的这套框架背后，是一套怎么样的关于 AI 和人类劳动关系的判断呢？

Kenny Zhang：我先说一下背景。我的理论只是基于现在 AI 的能力。我认为 AI 是一个非常非常强大的系统，我们现在可能只不过刚刚迈出了第一步，可能只用到了 AI 未来能力的 1%，甚至 0.1% 都不到。它未来的发展非常庞大。用我这个狭隘的认知，先说现在我认为 AI 能做的事情。

因为我特别喜欢历史，所以我们先回溯一下。以美国为代表，作为西方世界的代表，美国从 1985 年、1980 年代开始，制造业劳动人口大概有 2000 万左右。但是这个国家发展得非常快，工业化带动的需求飞速发展，国内劳动力远远不足以满足美国人的需求，所以有很大的通胀上行压力。

如果看 **80 年代**，通胀到了十几个点以上。他们没有办法，一定要寻找一个能降本增效、扩大生产规模、满足消费需求的地方。最开始先去了亚洲四小龙，韩国、台湾、日本、香港，还有一部分新加坡。但是这几个市场体量太小。

到了 **90 年代**，美国开始在中国做很多调研，看这个市场是不是可拓展。中国应该是 **2001 年**加入 WTO，正式进入自由贸易体系。这个时候美国把大量制造业迁移到了中国。

所以美国从 **1985 年**到现在，大概丢失了 700 万左右的制造业岗位，但是亚洲地区出现了大概 1.2 亿到 1.5 亿的制造业岗位。因为成本优化，转到中国的话，成本大概优化 **60%** 左右，又满足了美国的消费需求。

但你得有人去管这些产业链，对吧？那边多出了一亿多人，这边又多出了消费需求，你有更多人去卖、去管理下游产业链。到 **2010 年**，也就是这 **30 年**，大概增加了 4000 万白领。美国服务行业的就业一共是 1.3 亿，这 1.3 亿里边有 5000 万是最核心的白领。

2012 年，美国发现白领供给也不太够了，所以需要找第二个降本增效的出口。那个时候因为技术路线能提供的东西，它选择了云计算和软件。

但如果看那十年，其实没有一个特别革命性的变化。只说 **application 软件**，这些软件确实让美国不需要再增加白领，但也没有办法减少已有 **5000 万**白领的池子。云确实帮它省了一些成本，但已有 5000 万人的工资，以及为了维护这 5000 万人的成本还在继续增加。

SaaS 行业比较有趣，它只是把商业模式变了，这是非常成功的转型。在云之前，软件公司像我们小时候用 Windows XP，可能每三年买一次。因为本地部署，有很大的盗版问题，所以之前软件估值相对比较低。

到了 SaaS 层面，它变成了一个持续收费的公司。资本市场觉得这是最好的生意，因为每个月都得付钱。当它绑定了你的系统、拥有你的很多数据之后，就能去优化；软件每年还会提价。它进入你的公司之后，每年给你提价，你也没办法，因为你是在接受云软件提价和选择加人之间做选择，软件的 ROI 确实比加人要高。

这个持续了大概十几年，直到 AI 出现。在我的视角里，AI 是新时代对美国白领劳动力膨胀的一个解决方案。它已经解决了制造业那一块，制造业去了中国，中国市场很大，制造业效率一直在提升，也可以去东南亚。

知识劳动力这一块，AI 等于帮了美国大忙。我觉得明年开始，它会把很多劳动力的能力释放出来，终于能把这些人裁掉一部分，甚至一小部分到一大部分。**我觉得随着 AI 发展，它的实用性和稳定性如果得到接受，终于可以把这块成本降下来。**这可能是一个几年到十年的过程，慢慢白领就业增长的速度会下来。

到了这个时代，现在出现了一个新的数字经济。数字经济跟原来的经济没有差那么多，唯一的变量就是芯片要代替人。**以前中国制造业劳工，或者美国白领，要被芯片代替。**

但它的类比性很像中国以前工厂里的劳工：芯片是劳工，它们住在哪里？它们住在数据中心里，这是它们办公和生活的地方。谁来雇佣它们？就是采购这些芯片，以及给它们提供数据中心租金的超云和 Neocloud 公司。

所以英伟达为什么把它叫作 AI factory？因为它就像以前中国的工厂。你看它现在经历的很多困难、瓶颈，中国九十年代、两千年初都经历过：断水、断电、缺人，都是一样的。

网络设备，比如现在大家谈得比较多的光 and 连接设备，就相当于国内的高速公路和电信公司，这是让这些数字人交流沟通的地方。**最后，中国这边中国人要吃饭、喝酒、休息，芯片不休息，但它们需要更多电力。**所以这两个的可比性非常强。

我最后再重申一遍，这个是基于现在的 AI 能力。**现在已经有这么大一个平台式、板块式的知识劳动力出口，而且已经非常庞大，但我觉得我们还只是看到了 AI 发展的“小荷才露尖尖角”。**

曹卿云：对。谢谢 Kenny 今天跟我们的分享。我们从 1950 年代的大型机，一直聊到未来十年的基础设施将如何被 AI 定义。非常感谢今天参与节目，分享你的观点。

我们《硅谷坐标》这档节目最想做的事情，不是预测未来，而是搞清楚未来可能怎么发生，以及当下的我们站在哪里。我是曹卿云，这里是《硅谷坐标》，我们下期再见！

画面说明

图表标题为“数字劳工时代”，左侧对照 1990s 中国制造业的工人、工厂、电网、高速公路和食堂，右侧对照 2025+ AI Factory 的 GPU/CPU 数字劳工、数据中心、电力、光网络和冷却，中央箭头写着“芯片代替人”。另一张是 Kenny Zhang 访谈镜头，字幕提到芯片代替人。