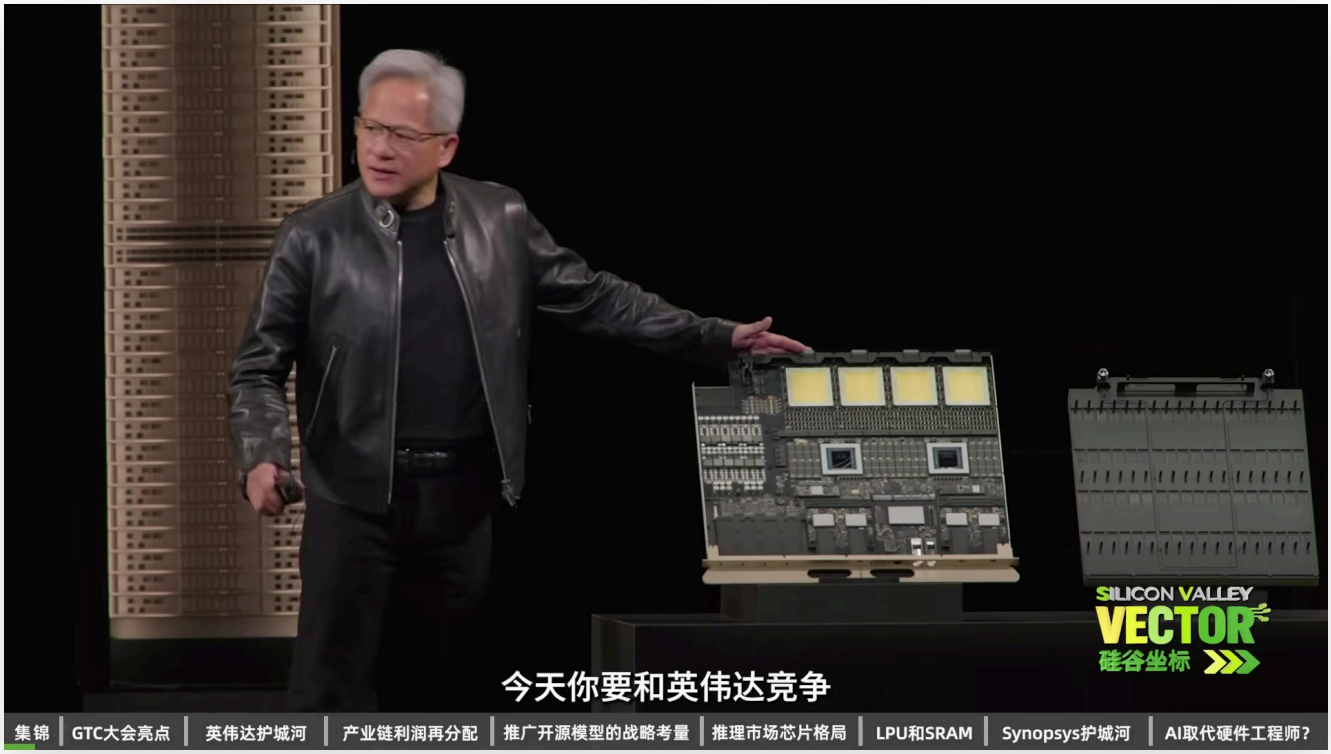
硅谷坐标 x CASPA副会长Jimmy Cheng:GTC回顾–AI时代英伟达的护城河 – 图文解说

> 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 Jimmy Cheng 结合 GTC 现场，解释英伟达如何从卖 GPU 走向 **AI Factory**、**软件生态**和**系统级方案**，并讨论 **Inference**、**LPU**、**SRAM**、**HBM**、**TPU**、**定制芯片**和 **Synopsys**。**内容的主线不是单颗芯片的性能，而是能源、基础设施、芯片、模型、应用和开发者生态共同构成的护城河。核心判断**是：AI **推理**时代会带来更强的专用化和多元化，但系统整合、软件兼容和已验证的工程能力仍是竞争壁垒。

01. GTC开场：AI Factory与Token竞争（00:03–01:30）



场景 1

本节主旨 GTC开场：AI Factory与Token竞争

完整内容

黄仁勋（GTC演讲片段）：Welcome to GTC. OpenClaw is number one; it's the most popular open-source project in the history of humanity. **It exceeded what Linux did in 30 years.**

黄仁勋（GTC演讲片段/中文转译）：**以后的每个 CEO 都会有一个 OpenClaw strategy，也就是 OpenClaw 的策略，就像以前每个 CEO 都会有一个 Linux 的策略一样。**

黄仁勋（GTC演讲片段）：This is your AI factory, this is your revenues. There's no question about that going forward. This is the throughput, this is the intelligence, better per watt for a given power of data center. The more throughput, the more tokens you could produce. **On this side is cost.** Notice, NVIDIA is the highest performance in the world.

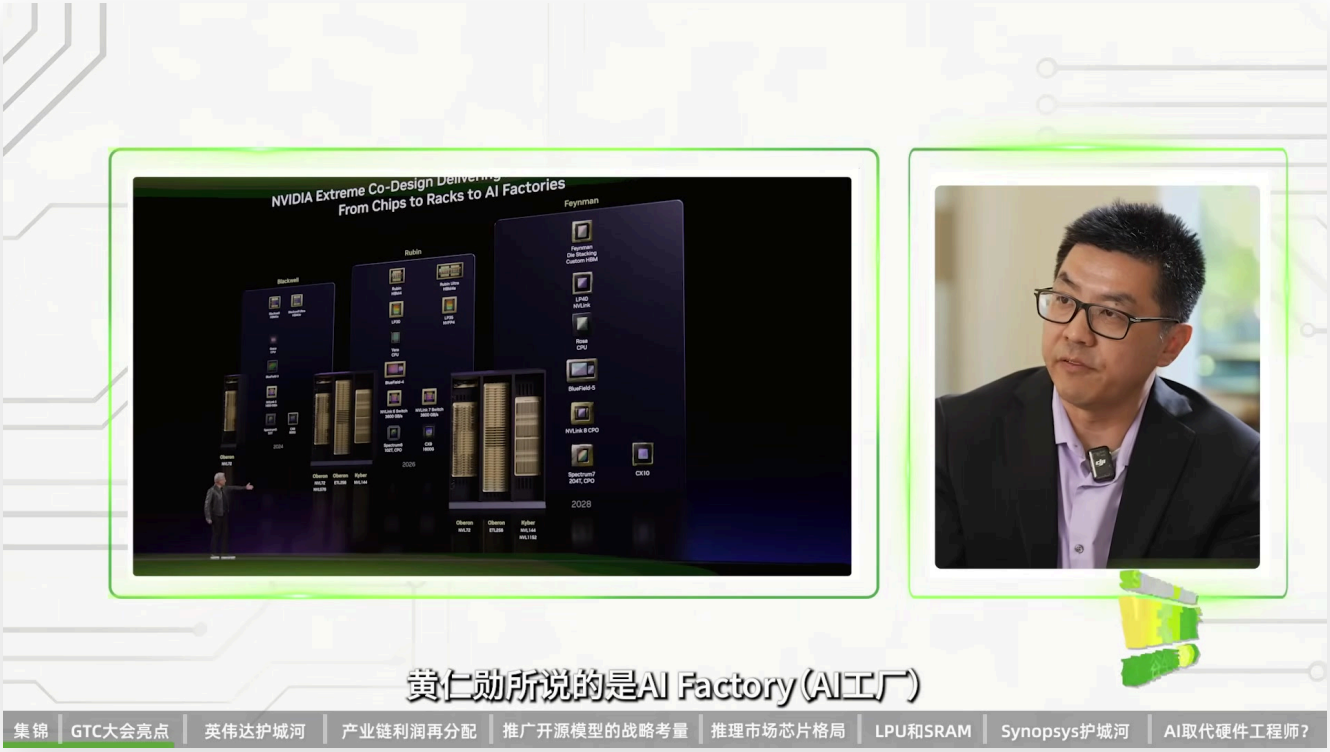
黄仁勋（GTC演讲片段/中文转译）：Nobody would be surprised by that。以前你出一个芯片，**10 倍的**速度就够了。今天你要和 NVIDIA 竞争，不是一个芯片 **10 倍的**速度就够了。**你必须要比它的 TCO 还低，你必须和它的软件兼容，所以系统化造成了非常大的护城河作用。**

黄仁勋（GTC演讲片段/中文转译）：开源模型的不断进步，会使芯片走向更加的通用化，还是更加的定制化？那开源化当然 drive the market，因为 demand 变大了。**Inference**，正如黄仁勋所说，这是一个非常大的 market。Market 越大，player 也越多，更多的小 startup 会进来，开发自己的 custom chip 路线。

画面说明

画面为黄仁勋在 GTC 舞台上讲解芯片与系统，右侧摆放两块芯片/板卡展示件；底部可见“今天你要和英伟达竞争”等中文字幕和节目章节条。

02. GTC三点感受：OpenClaw、Inference与Tokenomics（01:44–04:05）



场景 2

本节主旨 GTC三点感受：OpenClaw、Inference与Tokenomics

完整内容

曹卿云：Jimmy，欢迎来到《硅谷坐标》，和大家分享 GTC 的一些洞见。**你这次参加 GTC 最大的感受是什么?**

Jimmy Cheng：谢谢青云。对，我这次参加了 GTC，个人觉得这次 GTC 是一次很大的盛会，很多人用“春晚”来形容这次盛会，一点不为过。**黄仁勋一开场用了两个小时做 Keynote opening，我想给大家分享一下我的感觉。**

第一点，就是 OpenClaw。黄仁勋这次说，OpenClaw 是有史以来最伟大的软件之一，它的成长率已经超过了 Linux 和 Kubernetes，这两个都是在开源软件里成长非常快的软件。第二个，他还把 OpenClaw 说成是“Operating system of agentic AI”。**但是 OpenClaw 最大的问题就是 security，也就是安全和隐私。**针对这个方面，黄仁勋也宣布了 NemoClaw。NemoClaw 就是把 enterprise security、安全和隐私 build 在 OpenClaw 里面，这样适合企业来用。

他认为，以后的每个 CEO 都会有一个 OpenClaw strategy，也就是 OpenClaw 的策略，就像以前每个 CEO 都会有一个 Linux 的策略一样。**所以 OpenClaw 在整个 AGI 里是非常核心的一个地位。**

第二点就是 **Inference**。**Inference**，他提了不下 36 次。他于此也推出了新的产品 Vera Rubin，这是针对 Inference 设计的。同时，**LPU**，SongKwakSweet，也可以和 Vera Rubin 一起合作运用。

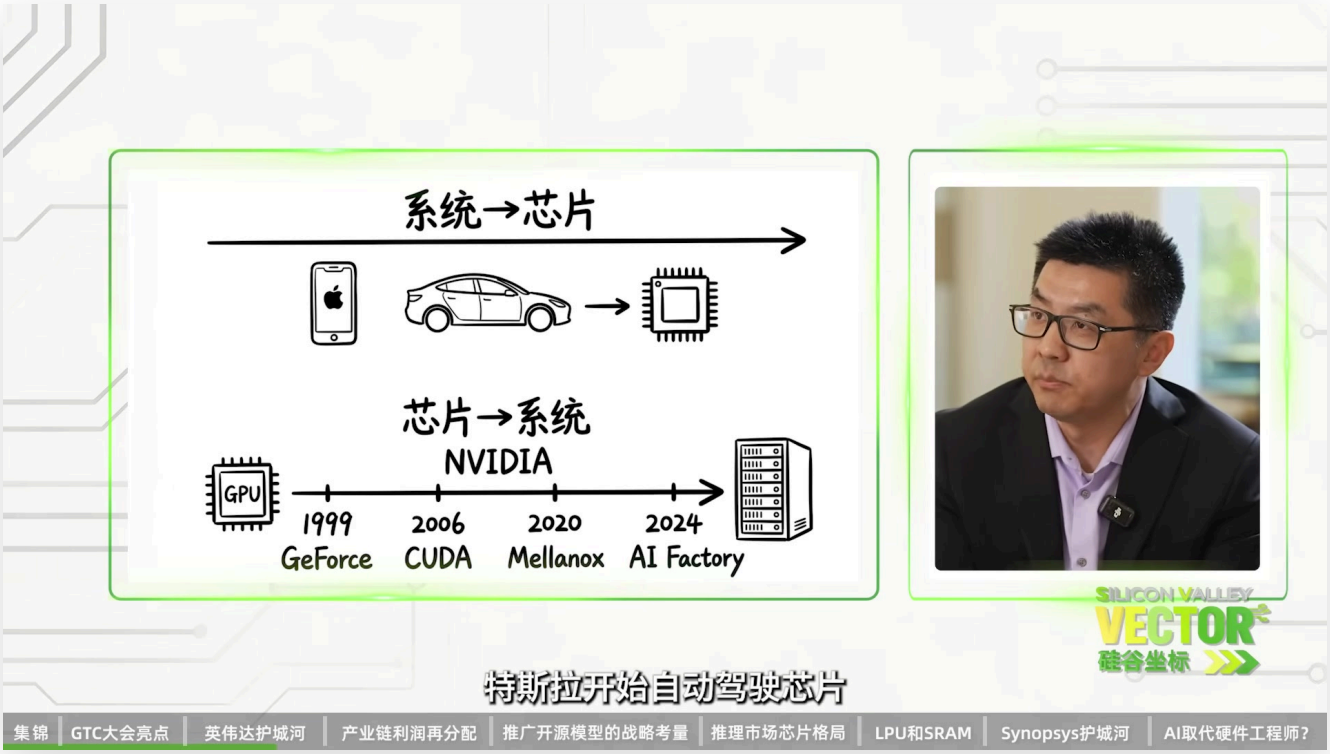
最后一点我想说的就是 tokenomics，这是一个新的 term，叫 token economy。相当于 token 是 currency of agentic AI。黄仁勋所说的是，AI factory 要和 AI factory 竞争。**任何的竞争对手，我们不是单纯地比 FLOPS、TOPS 了，而是 token per watt，所以这是一个升华。**

今年可以说是 OpenClaw 开启了一个新的时代，我们也看到了 OpenClaw 对于整个软件、硬件，甚至对于整个社会都产生了深远的影响。

画面说明

一张关键帧为 Jimmy Cheng 的近景访谈画面，底部字幕显示“**Inference（推理）**他提了不下36次”；另一张为 GTC 演讲系统图，标题写着“NVIDIA Extreme Co-Design Delivery: From Chips to Racks to AI Factories”，并列出 Blackwell、Rubin、Feynman 等阶段，右侧同时保留 Jimmy Cheng 画面；画面下方可见“黄仁勋所说的是**AI Factory**（AI工厂）”字幕和节目章节条。

03. AI Factory的五层与英伟达系统化路径（04:05–07:43）



场景 3

本节主旨 AI Factory的五层与英伟达系统化路径

完整内容

曹卿云：那我们今天也想重点谈一谈，就是 OpenClaw 的 **Agent** 时代对于半导体行业，特别是英伟达的一个影响。黄仁勋在他的主旨演讲里面提到，英伟达已经不再是一个卖芯片的公司，而是一个 **AI Factory**，也就是 AI 工厂。他把自己从一个 AI 芯片供应商的角色，提升到了一个系统解决方案供应商。你觉得这样的一个转变，对他的**护城河**来说是加强了还是减弱了？

Jimmy Cheng：哦，这是很好的一个问题。**AI Factory**，它是一个产生 token 的 data center。它有几个层面：第一层是 energy，第二层是 infrastructure，第三层是 chip，第四层是 model，第五层是 application。每一层都有它相对的东西。

Application 上是 Omniverse；model 层，这次它发布了 NeMoTron，以及其他的大 model，从 Anthropic 到 ChatGPT，这些都是 model 层面的。还有 infrastructure，InfiniBand 都是 infrastructure。Chip 当然是 GPU、CPU、**LPU**，这是非常重要的；**LPU** 特别是为 inference 设计的。Energy 包括 power delivery，以及 heat exchange 和怎么样处理 heat，也就是 heat management，等等。

曹卿云：所以你觉得，在这个 GPU 和 CUDA 的生态基础上，这些不同的 layers，包括存储的新架构、供电、高速网络通信、液冷，是不是这些新的架构都增强了它的护城河、增强了它的竞争力?

Jimmy Cheng：你说对了，它对这方面增强了非常大的竞争力。**我们看到两个趋势：第一个趋势就是芯片公司开始做系统；还有一个趋势就是系统公司做芯片。**

最好的 example 就是苹果和特斯拉。早期的苹果手机芯片不是自己研发的，而是三星研发的；到了第几代以后，苹果才逐渐把自己研发的芯片放在 iPhone 里面。同样，特斯拉开始的自动驾驶芯片是 Mobileye 做的，第二代也是 NVIDIA 做的，第三代才到自己。**所以这是一个从系统公司做芯片的趋势。**

NVIDIA 恰恰走的是另外一条方向，那就是从芯片公司走到系统化。黄仁勋也开了一个玩笑：他说以前买我们显卡的人都是父母，现在希望买显卡的是 developer 和 researcher 了。

当年 **1999 年**开发了 GeForce，它是第一批显卡；然后 **2006 年**出了 CUDA；之后 **2020 年**收购 Mellanox；它自研 Omniverse，**2024 年**出了大模型 NeMoTron。这些都是一步步走向系统化的步骤。它的系统已经很完善，所以它的护城河很坚固、很扎实。

所以，一般竞争对手如果想和 NVIDIA 竞争，现在难上加难。为什么？以前也许你出一个芯片，**10 倍的**速度就够了。今天你要和 NVIDIA 竞争，不是一个芯片 **10 倍的**速度就够了；你必须要比它的 TCO 还低，你必须和它的软件兼容。所以这个系统化造成了非常大的护城河作用。

曹卿云：在这种新的系统架构下，整个半导体产业的利润会怎么样重新分配? 议价权会往哪边走?

Jimmy Cheng：**现在英伟达不光是一个芯片公司，它也是个系统公司，它拥有全套的。**

画面说明

关键帧左侧是“系统→芯片”和“芯片→系统”的示意图：上方有手机、汽车和芯片，下方以 NVIDIA 为中心列出 1999 GeForce、2006 CUDA、2020 Mellanox、2024 **AI Factory**；右侧为 Jimmy Cheng 访谈画面。

04. 产业链利润与英伟达推广开源模型的考量（07:43–10:51）

场景 4

本节主旨 产业链利润与英伟达推广开源模型的考量

完整内容

Jimmy Cheng：系统。英伟达现在有 7 个芯片，很多芯片已经开始做不同的事，像 NVSwitch，相当于 networker。这个以前的供应商是 Cisco、HP 或者戴尔。英伟达也推出了 DGX，一个 blueprint。**所有这种 OEM 厂商都必须要 follow 它的 blueprint 了。**以前这些 OEM 还可以画些主板，但现在这个权利已经不在了，必须要 follow 英伟达 DGX 这种方案来做事，没有主动权。

它的 supplier 也处在一个很被动的局面，作为 supplier 利润很高，但是也存在 NVIDIA 做 double sourcing 的以后风险。所以在整个产业链里，英伟达占到绝大部分的利润。

曹卿云：我们知道，黄仁勋这一次在主旨演讲也讲到了，现在整个开源模型已经成为了第二大梯队，仅次于 OpenAI 的第二大生态玩家。他自己亲自主持了一个 panel，把主流的一些开源模型放在一起，然后做了一个研讨会。**他自己也下场，做了一个 NeMoTron 的开源模型，而且也在上面做了 OpenClaw。**自己不光是要做一个软硬件公司，也在做一个整个系统化的公司；同时他还在大力推广开源模型。你觉得这背后的考量是什么？

Jimmy Cheng：有几个原因。第一，他想增加 competitive pressure，在 Claude、Anthropic、OpenAI 上面。因为对他和英伟达最有益的是 model 公司不停地研发新的 model。每一个 model 的出现都会带动 hardware 的更新。**所以对他来说，他最不想看到的事情就是 model 公司不出新的 model，或者放慢脚步。**虽然 frontier model 大家都在努力往前进，但是他想把这个 cycle 变得更短、更快，这是他的一个原因。

还有一个原因，就是想把 GenAI 平民化。Open source 大家都知道是免费的。如果像 Salesforce 或 ServiceNow 这样的公司用 open model，那它唯一需要采买的是 NVIDIA 的 server，而不是 token。所以对他来说，可以促进更多 hardware 的销售。

第三个，还有一个地缘政治的原因。现在最领先的 open model，包括 DeepSeek 还有 Cohere，这些 model 是在中国研发的，这些 model 很可能是用非 NVIDIA 的 hardware。**因为 NVIDIA 不能卖到中国，所以它必须有一个强有力的 open model，来对抗中国的 Open Model，所以它就自己研发了一个 model。**

还有一个，当然大家都知道的，就是 NVIDIA 的策略。CUDA 已经有 700 万的开发者，它想通过 Open Source Model 更加增强它的 developer flywheel，把它自己的竞争力更加加强。**所以这几点就是为什么 NVIDIA 要开发 open-source model。**

画面说明

关键帧为主持人曹卿云的访谈近景；画面字幕显示“然后同时他还在大力的推广开源模型”，底部可见“推广开源模型的战略考量”章节标签。

05. 开源模型与定制化芯片（10:52–12:11）

场景 5

本节主旨 开源模型与定制化芯片

完整内容

曹卿云：开源模型的不断进步，会使芯片走向更加的通用化，还是更加的定制化?

Jimmy Cheng：这是很好的问题。我觉得开源模型会触动定制化芯片的发展。为什么这么说呢？如果把模型开源化以后，技术门槛会降低，做相应的 hardware 更容易了，因为模型是开源的，那些 weights 都能看见。**针对这些 weights，可以相对地做不同的 accelerator，所以这是各大公司想做的事情。**

第二个，如果一旦开源化了，像 ServiceNow、Salesforce，用得会很多。**Token 是免费的，但是他们还会买大量的芯片或大量的 server，从 NVIDIA 采购。**最后，很多公司也会在 capex 的逼迫下自己研发芯片，所以自己研发芯片也会走 custom chip 这条路线。

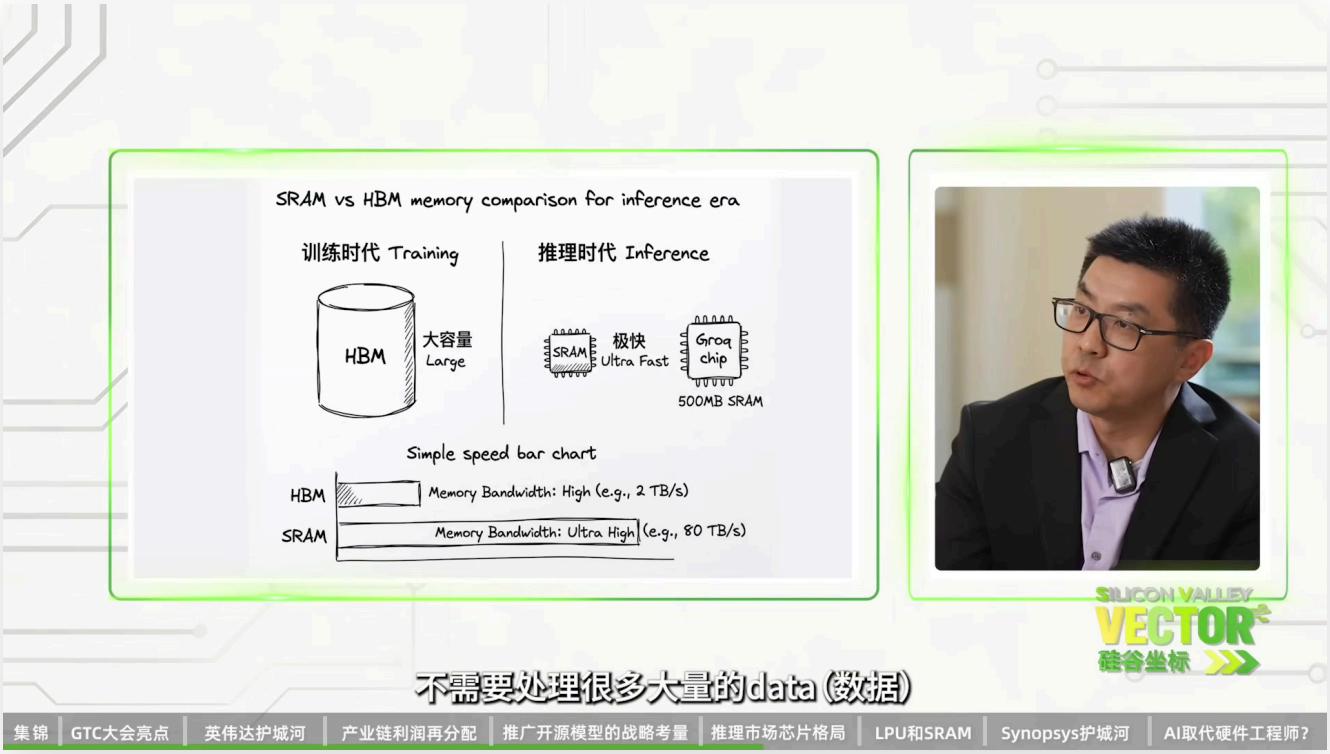
第三，开源化当然 drive the market，因为 demand 变大了。**Inference**，正如黄仁勋所说，这是一个非常大的 market。Market 越大，player 也越多，更多的小 startup 会进来，开发自己的 custom chip 路线。

曹卿云：刚刚也讲到今年是 agents 的元年，你觉得这会对底层的芯片架构设计提出什么样的新的要求?

画面说明

关键帧为 Jimmy Cheng 的访谈近景，底部字幕显示“但是他们还买量大的芯片”，下方章节条标出“推广开源模型的战略考量”和“**推理**市场芯片格局”。

06. SRAM-centric架构与3D芯片路线（12:12-13:51）



场景 6

本节主旨 SRAM-centric架构与3D芯片路线

完整内容

Jimmy Cheng：Agent 是**推理**时代的开始。**推理**这个时代其实 drive 一个 **SRAM**-centric 的 design philosophy。**SRAM**-centric 以前可能没怎么听过。

大家听的是 **HBM**。**HBM** 最近很火，因为很多公司生产 HBM 的都供不应求。但 SRAM 恰恰是我们所需要的一种 memory。SRAM 特别快，但是容量比较少，也就是量不大，但是很快。**Inference** 不需要处理很多大量的数据，但是要处理非常快的数据。**所以，这就是为什么它 drive SRAM-centric computing。**

我们为什么看到很多新兴的公司，像 Cerebras，最近 OpenAI 给它 10B dollar 的一个订单？它的 chip 非常大，像一个 iPad 那么大一样，上面很多都是 SRAM。还有包括 **LPU**，这次每个 chip 上有 500 megabytes 的 SRAM。像有个 LPX 是整个系统，NVIDIA 卖的是 192 个 **LPU**，192×500 MB，那是很大的 SRAM。这么多 SRAM 可以提高 inference 速度。

还有一个底层的方向，就是我们逐渐地走向 3D IC 的路线。**你可以看到，从 Hopper 那个时代，我们看到芯片还是单芯片；从 Blackwell 开始，芯片已经太大了，所以就分成两个芯片。**

曹卿云：**所以以后把 SRAM 累积在 GPU 上也是一个趋势。**我们也看到过去的三年，全球的 VC 投出了数百家的。

画面说明

关键帧一侧是“**SRAM** vs **HBM** memory comparison for inference era”图示：训练时代对应大容量 **HBM**，**推理**时代对应 ultra fast **SRAM** 和 Groq chip，并用条形图比较**内存**带宽；另一张为 Jimmy Cheng 近景。

07. LPU与GPU的分工（13:51–15:18）

场景 7

本节主旨 LPU与GPU的分工

完整内容

曹卿云：芯片的初创公司都想在**推理**市场上能够和英伟达抗衡。但我们也看到，去年 **2025 年 12 月**，英伟达 **200 亿美金**收购了 Groq 这家公司；而今年 **3 月份**，我们已经看到 Groq 被英伟达整合进自己的系统，发布了 **LPU** 芯片。你能跟我们讲讲 **LPU** 是怎么样工作的，然后它在英伟达的生态里有一个什么样的作用?

Jimmy Cheng：大家知道，GPU 强于 throughput，但弱于 latency；LPU 恰恰相反。如果对于 training，GPU 已经非常、也足够了；对 **Inference** 来说，LPU 在 latency 上很快的。我可以用一个比较生动的例子来比较 LPU 和 GPU。

LPU 好像一辆 taxi，出租车很快，但载的人很少。GPU 就像一个大巴，上面要坐 100 人，但是开得很慢，但一次可以载很多人。这就是 LPU 和 GPU 的相对比较。**LPU 和 GPU 的地位，不是说一个可以取代另外一个，它们都有需要，只是场景需要用得不一样。**

今年以来，**Inference** chip 一共投资了 **30 亿美金**，去年整一年投资是 **60 亿美金**。也就是说，三个月整个投资量已经达到了去年一半的水平，那就是增长率百分之**一百**。**所以 VC 是很看好 Inference 市场的。**

画面说明

关键帧是主持人与 Jimmy Cheng 的双人访谈远景，左下角标题条写着“**推理**市场与**LPU**的崛起”，画面字幕显示“GPU已经足够了”。

08. 物理AI的实时推理需求（15:22–16:18）

场景 8

本节主旨 物理AI的实时推理需求

完整内容

曹卿云：在整个新的架构下，你觉得 HBM 会起到更重要的作用吗?

Jimmy Cheng：Feynman 是 **2028 年**英伟达出的一款 GPU，这是一个 TSMC 1.6 nanometer 的 process。这个 process 是针对 physical AI，也就是所谓的机器人来设定的。**在这种场景下，什么样的东西最重要呢?** Real-time inference，及时的 inference，所以这个要求非常高。

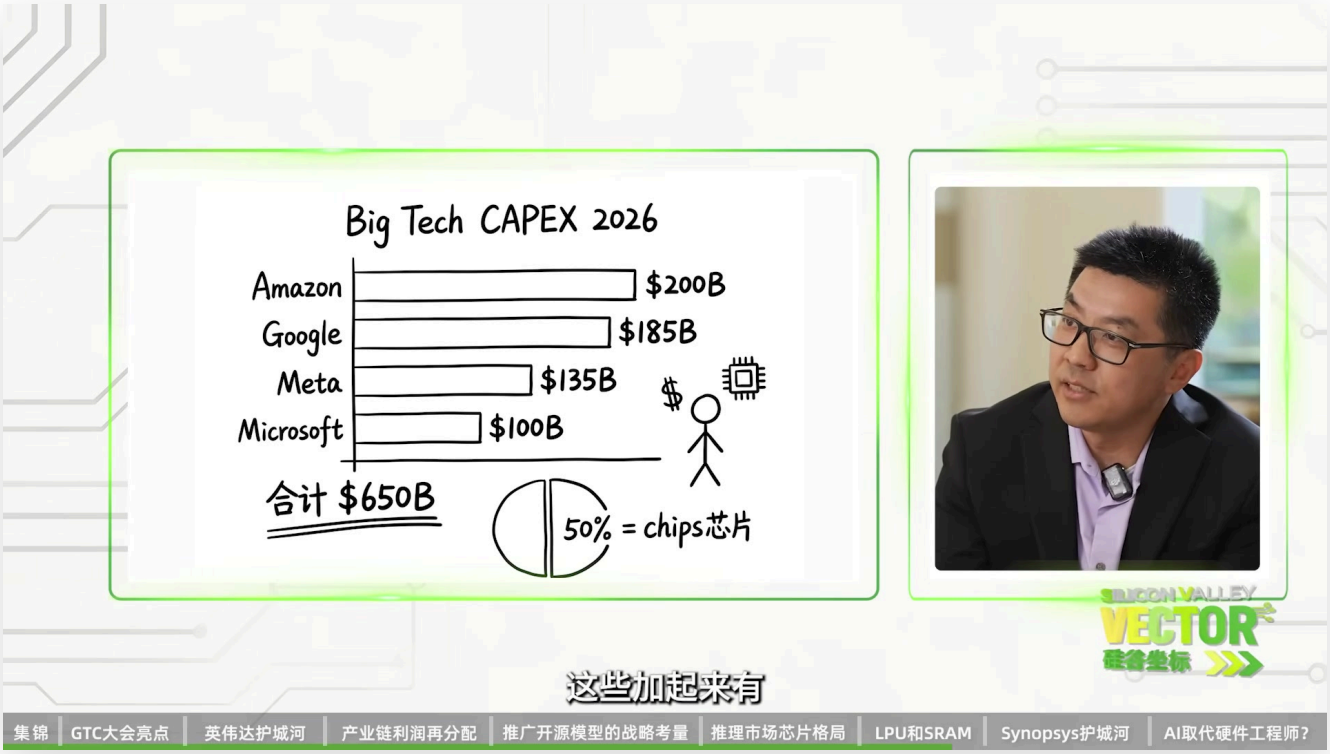
SRAM 的价值就体现出来了，所以 SRAM 会更多地被采用。而且很有可能，因为 SRAM 需要很多，普通的芯片已经不能满足，必须要用 3D 折叠的方式放在其他芯片上了。

曹卿云：这也是一个趋势。如果让你预测一下未来的**推理**芯片市场格局，我们有英伟达、谷歌的 TPU。

画面说明

关键帧为主持人与 Jimmy Cheng 的双人访谈远景，画面字幕显示“那普通的芯片已经不能满足”，底部章节条标出“LPU和SRAM”。

09. 自研芯片与多元化市场（16:18–19:57）



场景 9

本节主旨 自研芯片与多元化市场

完整内容

曹卿云：**然后我们有一些大厂做的自研芯片，还有 Cerebras 这样的初创公司。**你觉得这个未来的格局会是怎么样的?

Jimmy Cheng：首先我讲一下谷歌的 **TPU**。这款芯片是整个 camp 里面比较成功的一个芯片。为什么这么说呢? **TPU** 是针对 TensorFlow 定制的一款芯片。这次 Gemini 3.0 充分体现了它的价值。Gemini 3.0 所有的**推理**和 training 都是用它自己的芯片，完全没有用 GPU。这是一个很强的信号，对所有正在研发自己芯片的玩家来说，也就是说，英伟达的芯片也可以被战胜，谷歌已经证明了这一点。

所有其他的 player，像 AWS，AWS 也自己在研发自己的芯片。它有自己的 **Inference** 和 Training 芯片，**Inference** 芯片做得还不错，叫 Inferentia。AWS 也是一个大的 player。Meta 也在自己做自己的芯片，芯片叫 MTIA，这个还在研发，但是还在早期之中。**所以各大厂都在加速自己研发芯片的脚步，而不是停滞不前。**

大家如果 follow 所有大厂的 earning release，就会知道现在分析师进入 earning call，第一件事问的不是营收涨了多少、EPS 是多少，而是问你的 capex，这是一个非常重要的问题。Amazon 的 capex 是 200，Google 是 185，Meta 是 135，Microsoft 是 100。这些加起来有 650 billion 的 capex，一年；至少 **50%** 是花在芯片上的。650 是什么概念？就高于一个发展中国家的 GDP。

所以在这种环境下，大厂没有办法，必须要降低 cost，只有研发自己的芯片是一条路可以走。一开始可能付出的 R&D cost 会高，但一旦量产以后，每个 CPU、每个 GPU 的成本还是会比买 NVIDIA GPU 低很多。

第二个是 hardware 和 software co-design。这些大厂不光要自己做芯片，他们也自己有 model，有自己的 software stack。**用 model 和 chip 一起 design，可以达到最高的效率，所以大厂会继续往这方面投入。**

曹卿云：现在**推理**市场的芯片里，我们看到英伟达跟谷歌可以说是齐头并进的。未来你是怎么看他们核心竞争力的比较?

Jimmy Cheng：我个人觉得这个问题，其实决定于你相信未来是一个单一化的环境，英伟达一家独大，还是一个多元化的环境。**其实我更加倾向于多元化，因为 model 也是多元化，生产公司芯片也是多元化，不光是有英伟达，也有 NVIDIA，甚至 Meta 都会有自己的芯片，按自己的 model 来做。**以后包括 data center，也可能会是一个多元化的趋势。

最近有一种新的知识叫 UALink，UALink 是针对 NVLink 研制的，是一个 open format，NVLink 是一个 closed format。以后的趋势很可能是一个多元化的系统。

曹卿云：那谷歌和英伟达都会有自己的 role。接下来想问问 **Synopsys**，如何看待 AI 时代? **Synopsys** 的**护城河**是变强了还是变弱了?

画面说明

关键帧显示“Big Tech CAPEX 2026”柱状图：Amazon $200B、Google $185B、Meta $135B、Microsoft $100B，合计 $650B，并标注“**50%**=chips”；另一张为主持人曹卿云近景。

10. Synopsys的硅验证与AI设计工具（19:58–22:21）

场景 10

本节主旨 Synopsys的硅验证与AI设计工具

完整内容

曹卿云：因为我们听到一个声音是说，AI 工具制作的成本下降了，门槛下降了。以前只能有大厂才能做出来的一些复杂 IP，以后一些小厂也可以做出来了。**还有另一个声音说的是，随着 agents 的用量暴增，未来这些 agents 都需要 license，所以这是有利于 Synopsys 的。**您的观点呢?

Jimmy Cheng：那首先我来简单介绍一下我们公司。**Synopsys** 是全世界第 12 大的软件公司，也是第一大的 EDA 公司、第二大的 IP 公司，以及第一大的物理仿真公司。我们的 IP 排名第二，仅次于 Arm。AI 对所有生产力都有好的促进作用，所有研发芯片的公司都会 benefit，不只是 IP 的。

IP 这个 business 还是很特别的。其实很多人有错觉，觉得一个好的 IP 只要有一个 RTL code 能 work 就行了。但 IP 真正的价值不在于 RTL code work，而在于 silicon proven，也就是硅验证。这个 IP 被用过多少次，成功流片了多少次？很多 IP 被流片了 300 次、500 次，这不是 AI 一天两天可以把它的门槛降低的。**所以真正的核心价值在于 IP 是 silicon proven，而不是几行 RTL code。**

这次 Synopsys Converge，我们宣布了一个叫 L4 **Agent**ic Engineer 的工具。它仅次于 L5，L5 是最高级，full autonomous；L4 比它稍微低了一层。这个是什么概念呢？就是我们用 AI 来设计 AI，用我们的 AI agent 融合 NVIDIA 的 AI GPU，帮助各行各业的所有公司设计 AI 芯片。

我引用一下黄仁勋在我们 Converge 大会上说的话：他非常 excited，因为未来世界会有很多很多的 **Agent** 工程师，每一个工程师都会用我们的工具，那我们的工具会爆发性地成长。

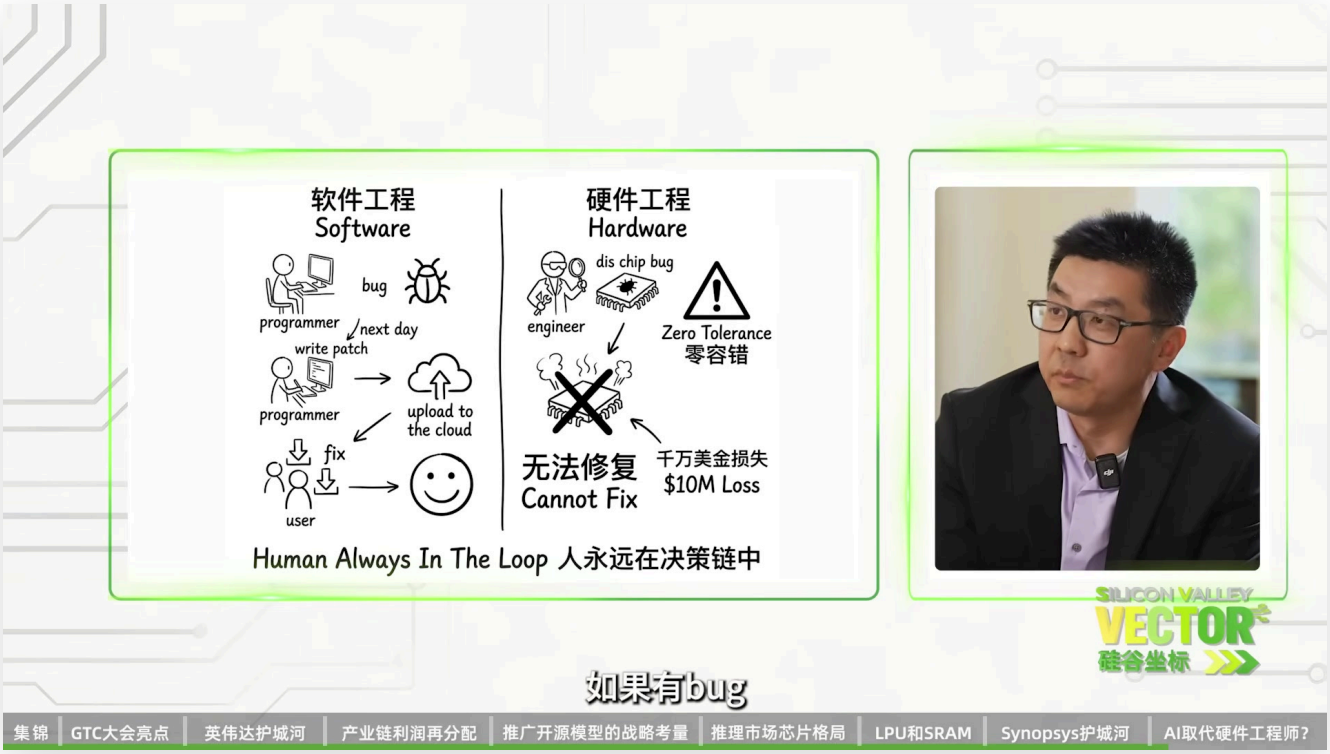
曹卿云：所以，你觉得 Synopsys 的**护城河**是变强了？

曹卿云：现在的程序员都面临被 AI 替代的危机，你觉得未来芯片设计师也会有这样的考虑吗？Fortune 一篇文章说，最近是 **40 年**以来。

画面说明

关键帧为 Jimmy Cheng 访谈近景，底部字幕显示“L5是最高级full autonomous”，画面下方保留节目章节条。

11. 硬件工程的Zero Tolerance（22:22–24:12）



场景 11

本节主旨 硬件工程的Zero Tolerance

完整内容

Jimmy Cheng：程序员就业市场最差的。四十年前，**1986 年**，那时 PC 还没有普及，Internet 还没有出现。那程序员面临的问题，是不是 hardware engineer 也会出现？其实程序员和 hardware engineer 本质上有很大的区别。

程序员写个程序，如果有 bug，第二天写个 patch 到 cloud 里面 deploy 了。但是 hardware engineer 是叫 zero tolerance 的行业。也就是说，如果我们的芯片有一个 bug，那这个芯片就不 work。研发一个芯片是很贵的一件事情，要一千万美金。**所以 zero tolerance 造成 human always in the loop，人的经验是不可替代的。**

第二点是我们这个行业现在还处处在人才荒的地步，其实很多 job 还没办法找到合适的人。

第二个，现在 schedule 也 push 得越来越紧。你也知道，Hopper 到 Blackwell 两年，Blackwell 到 Rubin 一年，加速了。**这意味着同样的人要做双倍的活，所以人才荒也是为什么短期之内 hardware engineer 不会像 software engineer 面临这种问题。**

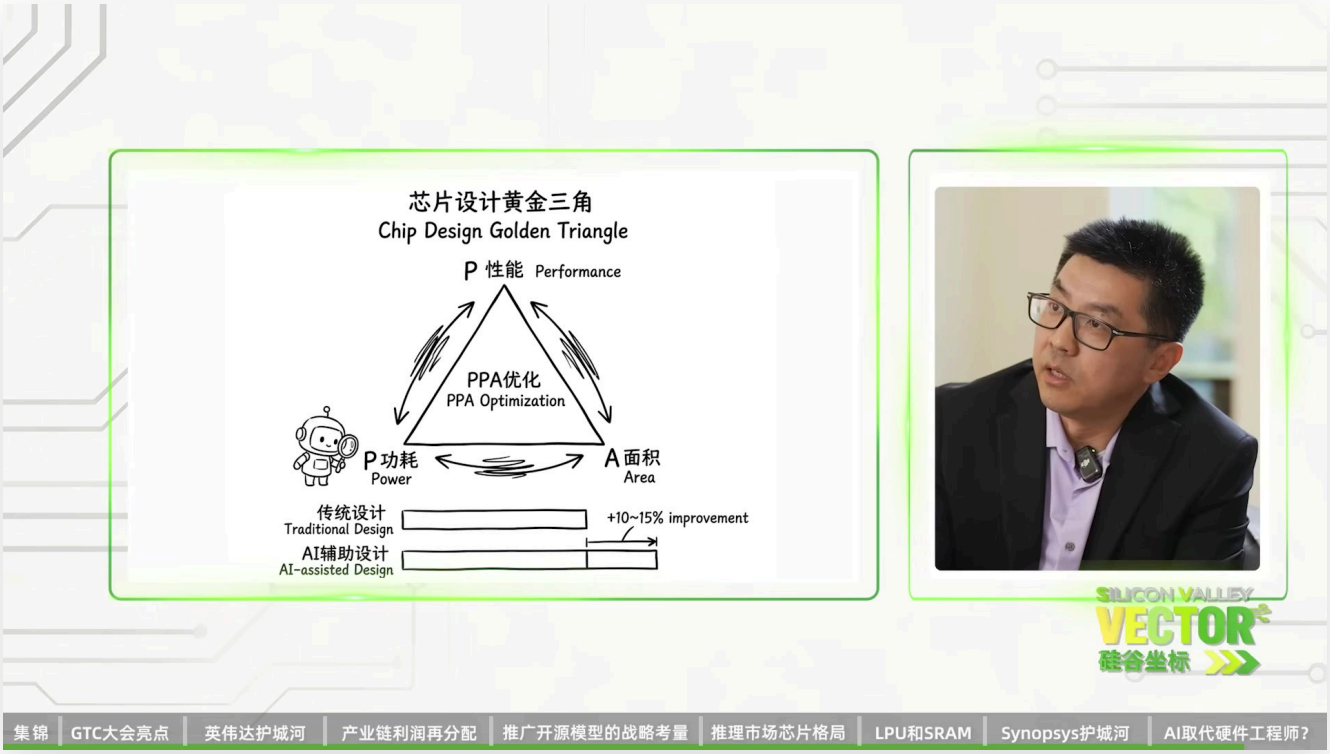
但长期来说，AI 是不是会取代一些 repetitive 的 job，或者是 manual 的 job？会的。但是我相信，hardware 工程师这些 job 会转移到其他更加有价值的 job。**因为有两个主要的原因：一个是 zero tolerance，还有一个是我们现在有人才荒的问题。**

未来人类设计的芯片和 AI 设计的芯片。

画面说明

关键帧左侧是软件工程与硬件工程对比图：软件一侧展示 bug、次日写 patch 并上传 cloud；硬件一侧标注“Zero Tolerance / 零容错”“无法修复”和“一千万美金损失”，底部写着“Human Always In The Loop”；**右侧为 Jimmy Cheng 访谈画面。**

12. AI辅助芯片设计与PPA优化（24:13–25:53）



场景 12

本节主旨 AI辅助芯片设计与PPA优化

完整内容

曹卿云：你觉得有些什么本质的区别？**你觉得未来会有一些设计出来的、性能很好的芯片，但是其实人类是读不懂的?**

Jimmy Cheng：对，我觉得 AI 设计的芯片和人类设计的芯片，我举一个简单的例子。**今天如果你对 AI 说：“我想设计一个两纳米的芯片，这是我的要求，是个 PLL，面积这么大。**”今天还没办法做到这点，还没聪明到这种程度。那以后会不会？我个人觉得我们不会等到这一天，至少短期内不会看到。为什么?

我们首先看一下 AI 在芯片这个行业的历史。早在七八年前，AI 就融入在芯片设计的方方面面了。芯片设计其实是很多步骤的：有 architectural design，也就是架构设计；前端包括 RTL、verification、synthesis，再加后端 sign-off。每个步骤其实 AI 已经 play a role，体现在 PPA 变好了。**PPA 就是 power、performance、area，这是我们衡量的标准。**

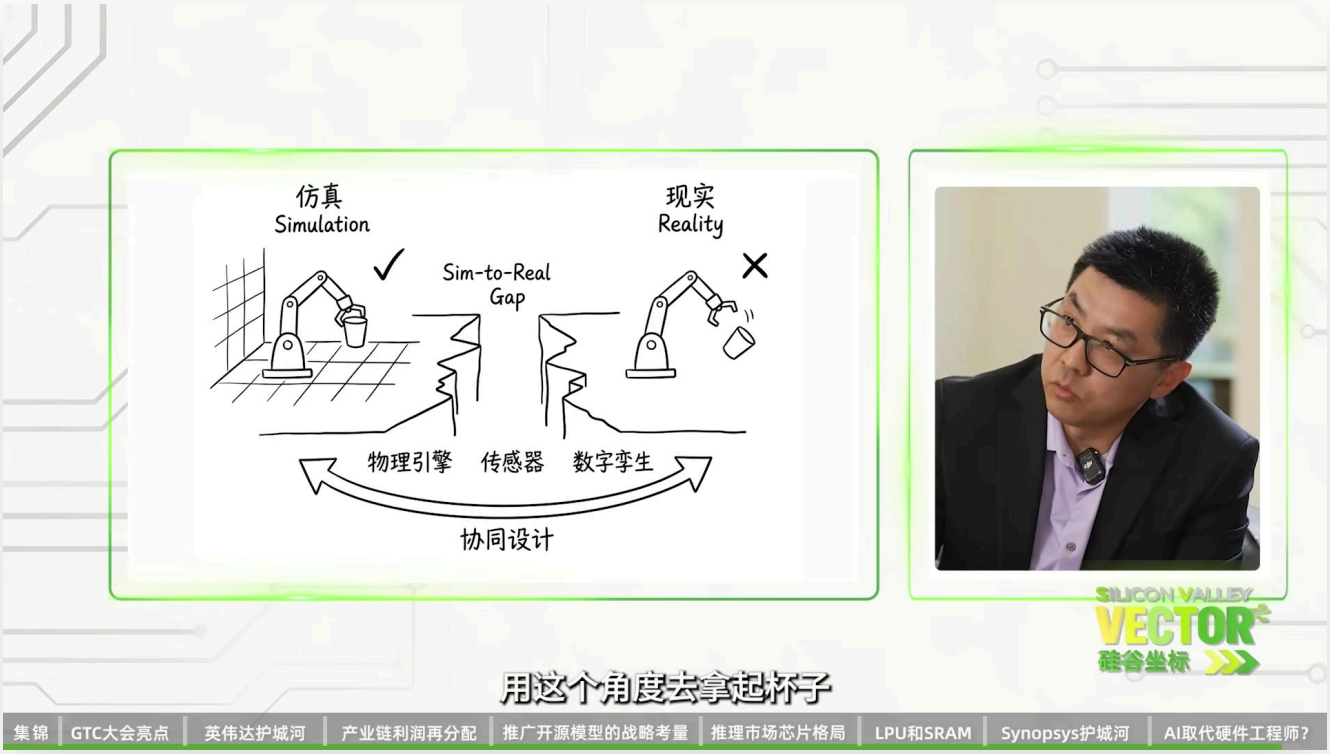
PPA 变好了，有可能达到 **10%** 或 **15%**，这已经达到了。现在我们正在第二个阶段，就是 agentic 的阶段，我们推出了 L4 **Agent**ic Engineer。

Jimmy Cheng：它会把很多 manual labor 取代掉，以后可能会达到 full autonomous。但是要达到黑箱操作，我觉得这个难度非常高。做黑箱，对有些行业可以，但对芯片设计是绝对不可以。为什么？**验证，或者 EDA 的 verification，是非常重要的。**

画面说明

关键帧显示“芯片设计黄金三角 / Chip Design Golden Triangle”：顶点为 P 性能、左下为 P 功耗、右下为 A 面积，中心为 PPA 优化；下方比较传统设计与 AI 辅助设计，并标注“+10~**15%** improvement”，右侧为 Jimmy Cheng 画面。

13. Physical AI的Sim-to-Real Gap（25:53-28:03）



场景 13

本节主旨 Physical AI的Sim-to-Real Gap

完整内容

曹卿云：这次参加完 GTC，你觉得现在整个半导体行业里，哪一个环节是最被忽视的?

Jimmy Cheng：我觉得有一个行业就是 physical AI 的仿真到现实的距离。Feynman 这个 chip 是 2028 年针对 physical AI 研发的，它其实就是一个 robot。那 robot 这个 system 是什么？它有 chip、有 model、sensor，还有 actuator，也就是传感器和执行器。

为什么 sim-to-real gap 是个大问题？**我举个例子，如果你 simulate 说用这么大的力度、用这个角度去拿起杯子，结果机器人用这个角度、用这么大力度还拿不起来，那就是有一个 gap。**

这方面有几个 challenge。第一个就是 real-time 的 physics engine。**如果一个球掉下来，我能知道它几秒钟掉下来，这个引擎就必须了解 gravity，也就是重力。**

第二个是 real-time 的 sensor processing。这也是非常重要的。Sensor 像 camera、LiDAR，要做 real-time 的 processing。**这个 latency 很重要，还要 deterministic，也就是确定性。**三秒完成就是三秒。为什么？如果你晚一秒，也许这个杯子已经掉到地上了。三秒钟完成这个运算是很重要的。

第三个是 real-time 的 digital twin，也就是数字孪生。**如果一个杯子要掉下来之前，你做一百次模拟，把各种方方面面都模拟过了，这个很重要。**但是这方面也没有办法及时完成。

很多 challenge 需要以后我们 hardware 和 software co-design，协同联合在一起，我们就可以看到一个很好的未来。我是一个 science fiction 的 fan，未来已经离我们很近了。我们再努力一把，把 sim-to-real gap close，就更加能达到未来。

非常感谢今天你的时间，感谢你的分享。

Jimmy Cheng：谢谢！

画面说明

关键帧显示“仿真 / Simulation”与“现实 / Reality”两侧的机械臂和杯子，中间标注“Sim-to-Real Gap”，下方列出物理引擎、传感器、数字孪生和协同设计；另一张为 Jimmy Cheng 访谈近景。