

An Interview with CoreWeave Executives: AI Demand Seems to 'Intensify' Every Day

CoreWeave 高管访谈：AI 需求似乎每天都在“加剧”

Key Context spoke with CoreWeave co-founder Brannin McBee and VP of Corporate Development Nick Robbins to discuss the state of AI demand.

Key Context 与 CoreWeave 联合创始人 Brannin McBee 以及企业发展副总裁 Nick Robbins 进行了访谈，探讨了 AI 需求的现状。

TAE KIM
JUN 17, 2026 • PAID

CoreWeave is regarded as the innovative early market leader of the neocloud sector.

CoreWeave 被视为新型云领域具有创新性的早期市场领导者。

It is the only vendor to earn a top Platinum-tier rating from AI research firm SemiAnalysis. Founded in 2017, the company provides large-scale GPU capacity to startups and enterprises.

它是唯一一家获得 AI 研究公司 SemiAnalysis 顶级白金评级的供应商。该公司成立于 2017 年，为初创企业和大型企业提供大规模 GPU 容量。

Key Context recently spoke with CoreWeave co-founder and Chief Development Officer Brannin McBee and VP of Corporate Development and Investor Relations Nick Robbins to discuss the state of AI demand and the neocloud market.

Key 近期采访了 CoreWeave 联合创始人兼首席发展官 Brannin McBee，以及企业发展与投资者关系副总裁 Nick Robbins，共同探讨 AI 需求和新兴云市场的现状。

Here are lightly edited highlights from our conversation:

以下是经我们整理的对话精彩片段：

Tae: When did the agentic AI demand wave take off?

Tae: 智能体 AI 需求浪潮是何时兴起的？

Brannin: We saw the real beginnings of that in Q4 as we were having the engineering conversations with our clients of what they were expecting to bring to market in Q1.

布兰宁：我们在第四季度就看到了这一切的开端，当时我们与客户就其预计在第一季度推向市场的内容进行工程层面的探讨。

That’s a really important lens that we consistently have with our clients. It’s a deep interconnected engineering relationship. That allows us to be ahead of where everything is going instead of reactive.

这是我们要始终与客户保持的一个非常重要视角。这是一种深度互联的工程关系。这使我们能够走在趋势的前面，而不是被动应对。

I would say from a product perspective for the AI market, Q1 was just this massive inflection in inference and AI consumption that only continues to accelerate.

我想说，从 AI 市场的产品角度来看，第一季度在推理和 AI 消费方面出现了巨大的拐点，而这种加速态势仍在持续。

Tae: What is the state of AI demand right now? Would you say there is no letup in the last few weeks versus months ago?

Tae: 目前 AI 需求状况如何？您会说最近几周相比几个月前没有任何放缓迹象吗？

Nick: It finds new ways to intensify seemingly every day.

Nick: 这种需求似乎每天都在找到新的方式加剧。

Tae: Tell us about the rising trend of CPUs versus GPUs in the agentic AI wave. Will you install rows of Vera CPU racks next to Nvidia GPU servers?

Tae: 请谈谈在代理式 AI 浪潮中 CPU 与 GPU 的上升趋势。你们会在 Nvidia GPU 服务器旁边安装一排排的 Vera CPU 机架吗？

Brannin: CoreWeave has been running CPUs since 2023. We’ve had that full cloud product consistently. It’s not, are we beginning to add CPU? It’s more what is the client asking for? And is it increasing on a relative basis? Absolutely, yes.

Brannin: CoreWeave 自 2023 年以来一直在运行 CPU。我们一直拥有完整的云产品。问题不在于我们是否开始添加 CPU，而在于客户在要求什么？相对而言，需求是否在增加？绝对是肯定的。

Storage has increased on a relative basis to prior generations as well, as agentic and reasoning have really taken off within the models.

随着代理式 AI 和推理能力在模型中真正兴起，存储相对于前几代也有所增加。

I see that trend continuing.

我预计这一趋势将持续下去。

Nick: The answer to your question is yes. You’re absolutely going to see a bunch of Vera CPUs crammed next to a bunch of Vera Rubin servers. Last year, we effectively fundamentally re-architected our base case data center design to leave room for a lot more storage and a lot more CPUs to live next to the GPUs.

Nick: 你的问题的答案是肯定的。你绝对会看到大量的 Vera CPU 与大量的 Vera Rubin 服务器紧挨着放置。去年，我们从根本上有效地重新架构了我们的基础数据中心设计，为更多的存储和更多的 CPU 留出了空间，让它们能够紧邻 GPU 运行。

The reason we did that is informed by this unique place we sit in this ecosystem. We’re the only independent cloud that serves all the most sophisticated users of this technology. Like there’s no other independent AI cloud that can say, Anthropic and OpenAI and Meta and Google and Microsoft and Nvidia and so on are customers.

我们这样做的原因，源于我们在这一个生态系统中所占据的独特位置。我们是唯一一家服务于该技术所有最尖端用户的独立云服务商。没有其他独立的 AI 云服务商能够声称，Anthropic、OpenAI、Meta、Google、Microsoft 和 Nvidia 等都是他们的客户。

That’s created this helpful flywheel for our business or positive reinforcement loop where we understand where our customers are taking the technology and we plan accordingly.

这为我们的业务创造了一个有益的飞轮效应，或者说正向反馈循环，让我们能够了解客户将技术引向何方，并据此进行规划。

Key Context by Tae Kim is a reader-supported publication. To receive new posts and support my work, consider becoming a free or paid subscriber.

Upgrade to paid

Tae: Are you primarily going to use Nvidia Vera CPUs going forward?

Tae: 你们未来是否将主要使用 Nvidia Vera CPU？

Nick: It depends on the workload. We’re customer led in what we do. We definitely expect to be an early and meaningful adopter of Vera CPUs. We’ve already disclosed that. Today, the fleet is actually mostly AMD at the moment, but that probably shifts over time based on customer demand and there’s a ton of interest for Vera CPU.

Nick: 这取决于工作负载。我们的行动是由客户主导的。我们确实期望成为 Vera CPU 的早期且重要的采用者。我们已经披露了这一点。目前，我们的机群实际上主要由 AMD 构成，但随着时间的推移，这可能会根据客户需求而变化，而且大家对 Vera CPU 有着浓厚的兴趣。

Brannin: That’s a good reminder just on how our contracts work. As you know, 98 plus percent of our revenue is contractually driven. We aren’t guessing at what clients want for the infrastructure. They’re telling us exactly what builds they’re looking for. It is like client led. They are defining what it is that we are building.

Brannin: 这是一个很好的提醒，可以借此谈谈我们的合同运作方式。如你所知，我们 98% 以上的收入是由合同驱动的。我们不需要猜测客户需要什么样的基础设施。他们会确切地告诉我们他们在寻找什么样的配置。这一切都是以客户为主导的。正是他们在定义我们要建设什么。

Tae: Talk about the competitive landscape. How do you go to market versus neoclouds like SpaceX, Nebius, Oracle versus hyperscalers like Azure, AWS and Google?

Tae: 谈谈竞争格局。在市场推广方面，你们是如何应对像 SpaceX、Nebius、Oracle 这样的 neocloud（新云厂商），以及像 Azure、AWS 和 Google 这样的超大规模云厂商的？

Brannin: The perspective I like to take on differentiation sits with the third-party verification that’s out there. Nine out of the ten top AI labs, excluding China, use our platform. SemiAnalysis consistently ranks us singular in our performance. I don’t think it’s out of our friendship with Jensen that we get allocated GPUs the way we do.

Brannin: 关于差异化，我倾向于从第三方验证的角度来看。除中国外，全球排名前十的 AI 实验室中有九家在使用我们的平台。SemiAnalysis 始终在性能方面将我们列为独一无二的存在。我认为我们之所以能以这种方式获得 GPU 分配，并非是因为我们与黄仁勋的交情。

That’s a supplier who has deep confidence in our execution track record and engineering capabilities to be the best representation of his product on the planet.

那是一位供应商，他对我们的执行记录和工程能力充满信心，认为我们是他的产品在全球范围内的最佳代表。

Nick: We win with the hyperscaler because we’re really good at execution. We can stand these things up incredibly quickly and they run very well. We’re winning with the research lab because we deliver the most performant version of the technology and the efficiency on a per token basis.

Nick: 我们在超大规模云服务商方面胜出，是因为我们非常擅长执行。我们能够极其迅速地部署这些设施，而且运行状况非常好。我们在研究实验室方面胜出，是因为我们提供了性能最佳的技术，并且在每个 Token 的基础上实现了最高的效率。

We’re winning with the enterprise because of, yes, the infrastructure works well and we built a really nice orchestration layer that’s best in class and that’s where platinum ranking and all that stuff comes in.

我们在企业客户方面胜出，是因为是的，基础设施运行良好，并且我们构建了一流的编排层，这正是白金评级等成就的来源。

But increasingly importantly, it’s because we’ve built, certainly amongst the AI clouds, the most sophisticated top of the layer across inference and development tools that empowers enterprises to productionize AI.

但日益重要的是，因为在所有 AI 云中，我们构建了最顶尖的推理和开发工具层，赋能企业将 AI 投入生产。

What that means is we are building and delivering products that ultimately help enterprises who are less sophisticated take data and turn it into models, turn it into agents that they can then run internally and we can cross-sell CoreWeave cloud into.

这意味着我们正在构建和交付产品，最终帮助那些不太成熟的企业获取数据并将其转化为模型、转化为智能体，供其内部运行，从而让我们能交叉销售 CoreWeave 云服务。

Tae: What are the current bottlenecks? Data center powered shells? GPUs? Electricians?

Tae: 当前的瓶颈是什么？数据中心配备电力外壳的设施？GPU？还是电工？

Brannin: Powered shells. It’s the components within powered shell. You’re absolutely correct to highlight electricians. It is a complicated space, but I think what’s important there is we have 49 of these sites online and running. We’re not just putting hopes and prayers into one or two sites. We’ve done this 49 times.

Brannin: 配备电力外壳的设施。是这些设施内部的组件。你强调电工这一点完全正确。这是一个复杂的领域，但我认为重要的是，我们要有 49 个这样的站点已上线并运行。我们并不是只把希望寄托在一两个站点上。我们已经这样做了 49 次。

That’s an immense amount of execution track record.

那是一份非常丰富的执行履历。

That’s an immense amount of knowledge that we have built on how to deal with supply chain problems and who are the right vendors and the wrong vendors to be working with within that supply chain.

那是我们在如何应对供应链问题、以及在供应链中与哪些供应商合作才是正确或错误的选择等方面积累的深厚知识。

(Editor’s note: Powered shells are the data center buildings and do not include the actual computing server hardware.)

(编者注：Powered shells 指的是数据中心建筑，不包括实际的计算服务器硬件。)

Tae: What can you tell us about HBM memory costs and shortages? How are you’re dealing with that? Do customers have to pay for the price hikes?

Tae: 关于 HBM 内存的成本和短缺，您能告诉我们些什么？你们是如何应对的？客户是否需要承担价格上涨的费用？

Nick: The answer to that is yes. The way we designed our business model is to lock in what we’re charging for the GPUs at the same time that we’re signing purchase orders for what we will be charged for the GPUs. Or servers more broadly, which obviously includes an HBM component.

Nick: 答案是肯定的。我们设计商业模式的方式是，在我们签署购买订单以锁定我们要为 GPU 支付的费用的同时，锁定我们要向客户收取的 GPU 费用。或者更广泛地说，是服务器费用，其中显然包含了 HBM 组件的成本。

And that is our way of insulating ourselves from the day-to-day price moves.

这就是我们保护自己免受日常价格波动影响的方式。

If price goes higher on our component price on the next deal, we make that into what we’re comfortable charging the customer, and therefore we protect our margin. We’re very well protected from it in our ability to pass through those costs onto the customer. It’s something that we monitor very closely.

如果在下一笔交易中我们的组件成本上涨，我们会将其调整为我们认为可以向客户收取的合适价格，从而保护我们的利润率。我们能够将那些成本转嫁给客户，因此在这一点上我们受到了很好的保护。这是我们密切关注的事情。

Access to components is not the biggest bottleneck today. Powered shell is. But there may be points in time in the future in which that answer flip flops.

如今，获取组件并非最大的瓶颈，带外壳的供电设施才是。但在未来某些时刻，这一答案可能会发生逆转。

Tae: How do you expect the Vera Rubin ramp to go? How is it going to go in the second half of this year?

Tae: 你预计 Vera Rubin 的 ramps 将如何进行？今年下半年情况会如何？

Nick: We obviously were the first in the world to bring up and fully validate a VR [Vera Rubin] cabinet. As we were with GB200, GB300 last year, I expect VRs to start to show up later this year.

Nick: 我们显然是全球首家启用并全面验证 VR [Vera Rubin] 机柜的公司。就像去年我们在 GB200、GB300 上做的一样，我预计 VR 将在今年晚些时候开始出现。

I expect that heavy, heavy ramping to be throughout 2027 in the same way that GB started showing up in 2025, but the heavy, heavy ramp was really throughout 26. Like you had plenty stood up towards the end of last year, but this year has been really ramping GB at scale.

我预计大规模 ramping 将在整个 2027 年持续，就像 GB 在 2025 年开始出现，但大规模 ramping 实际上贯穿了整个 2026 年。就像你在去年年底已经部署了很多，但今年才是真正大规模 ramping GB 的一年。

I expect there to be a pretty similar pattern for VR over the course of the next 12 to 18 months.

我预计在接下来的 12 到 18 个月里，VR 将呈现非常相似的模式。

Tae: On the Nvidia vs. ASICs debate. What are the customers saying now? In the past I heard from startups who say they are sticking with Nvidia versus Amazon because Nvidia is the most optimized, most reliable, battle tested. Is it still the case?

Tae: 关于英伟达与 ASIC 的争论。客户现在怎么说？过去我听到过一些初创公司表示，他们坚持选而不是亚马逊，因为英伟达最优化、最可靠、最可靠。现在还是这样吗？

