OpenAI Jalapeño: Better Than Nvidia Blackwell

OpenAI Jalapeño：优于 NVIDIA Blackwell

OpenAI's self-designed ASIC compared with Rubin, Jalapeño's TCO, throughput per MW, and spicy deets

OpenAI 自研 ASIC 与 Rubin 对比、Jalapeño 的 TCO、每 MW 吞吐量，以及更多劲爆细节

BRYAN SHAN, MYRON XIE, JORDAN NANOS, AND 3 OTHERS

BRYAN SHAN、MYRON XIE、JORDAN NANOS 及其他 3 人

AUG 25, 2026 • PAID • 付费内容

OpenAI has spent the past couple years quietly building "Jalapeño," an inference chip just announced at Hot Chips. Rumors of a successful tapeout had been swirling for a while. But now we have details. OpenAI invited us to look at their chip, go to their labs to check out how real it is, and benchmark it with our InferenceX suite.

OpenAI 过去几年一直在低调打造“Jalapeño”，这款推理芯片刚刚在 Hot Chips 上发布。此前关于流片成功的传闻已流传一段时间。但现在我们有了具体细节。OpenAI 邀请我们参观了他们的芯片，前往其实验室实地考察其真实程度，并用我们的 InferenceX 套件对其进行了基准测试。

In June, OpenAI unveiled the chip program in partnership with Broadcom, built from a blank slate exclusively for LLM inference. Design work began in the middle of 2024, going from initial team hiring to manufacturing tape-out in ~16 months, an extremely fast ASIC development cycle.

6 月，OpenAI 公布了与博通（Broadcom）合作的芯片项目，该芯片从零开始设计，专为 LLM 推理打造。设计工作始于 2024 年年中，从初始团队组建到制造流片仅用时 16 个月，这是一个极快的 ASIC 开发周期。

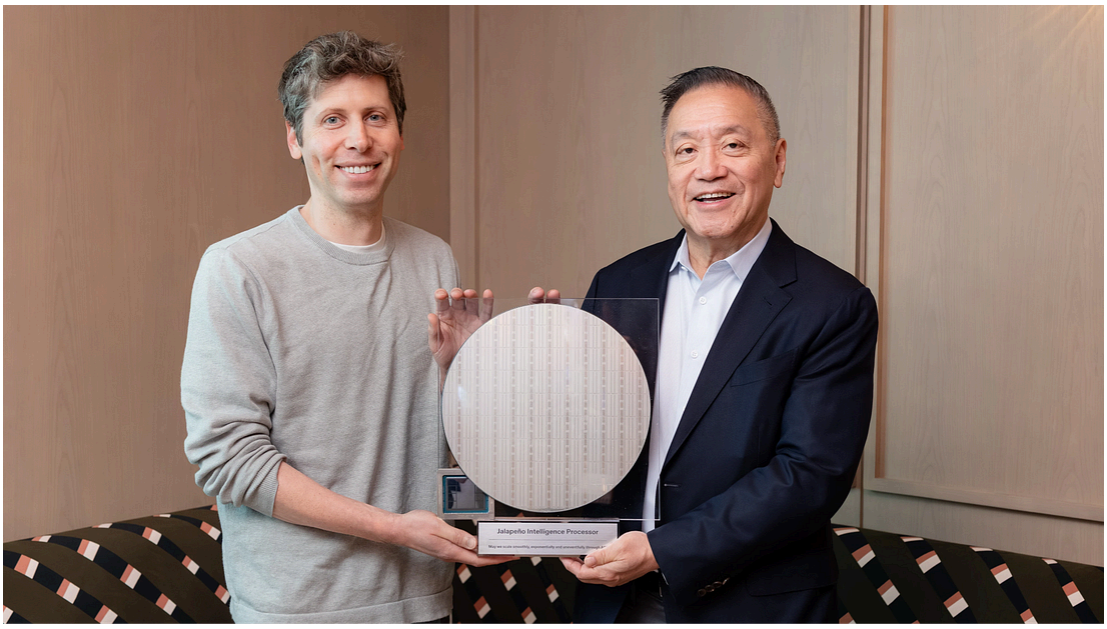
In general first generation chips are not competitive, but OpenAI bucks the trend by being industry leading and beating every Nvidia, AMD, and Google chip we have been able to test on multiple top open source models. OpenAI does this with extreme

hardware software codesign. Surprisingly, OpenAI is not over specialization on any specific part of model inference, but instead by focusing on being a general chip that delivers high performance in all scenarios.

第一代芯片通常不具备竞争力，但 OpenAI 打破了这一惯例，凭借行业领先的表现，在多个顶级开源模型上击败了我们所能测试的所有 NVIDIA、AMD 和 Google 芯片。OpenAI 通过极致的软硬件协同设计实现了这一点。令人意外的是，OpenAI 并未针对模型推理的某一特定环节进行过度定制，而是专注于打造一款通用芯片，在所有场景下都能提供高性能。

In this article, we will go into architectural details, software details and performance results for Jalapeño on InferenceX.

在本文中，我们将深入探讨 Jalapeño 在 InferenceX 上的架构细节、软件细节及性能表现。



Source: OpenAI OpenAI

A generalized inference chip

一款通用推理芯片

Everyone says that OpenAI's chip is specialized for OpenAI models, but that's wrong, OpenAI made a generalized chip for AI inference.

所有人都说 OpenAI 的芯片是为 OpenAI 模型专门定制的，但这是错误的——OpenAI 打造的是面向 AI 推理的通用芯片。

The timelines are insane. It shows that claims that use of AI is being used to accelerate chip design are real. Regardless of the quick timelines,Open AI spent a bunch of money, made pragmatic design decisions and their team is cracked, so this comes as no surprise.

时间线令人难以置信。这表明“AI 被用于加速芯片设计”的说法是真实的。抛开快速的时间线不谈，OpenAI 投入了大量资金，做出了务实的设计决策，而且其团队实力超群，因此这并不令人意外。

Just looking at the specs, it is an immediate contender:

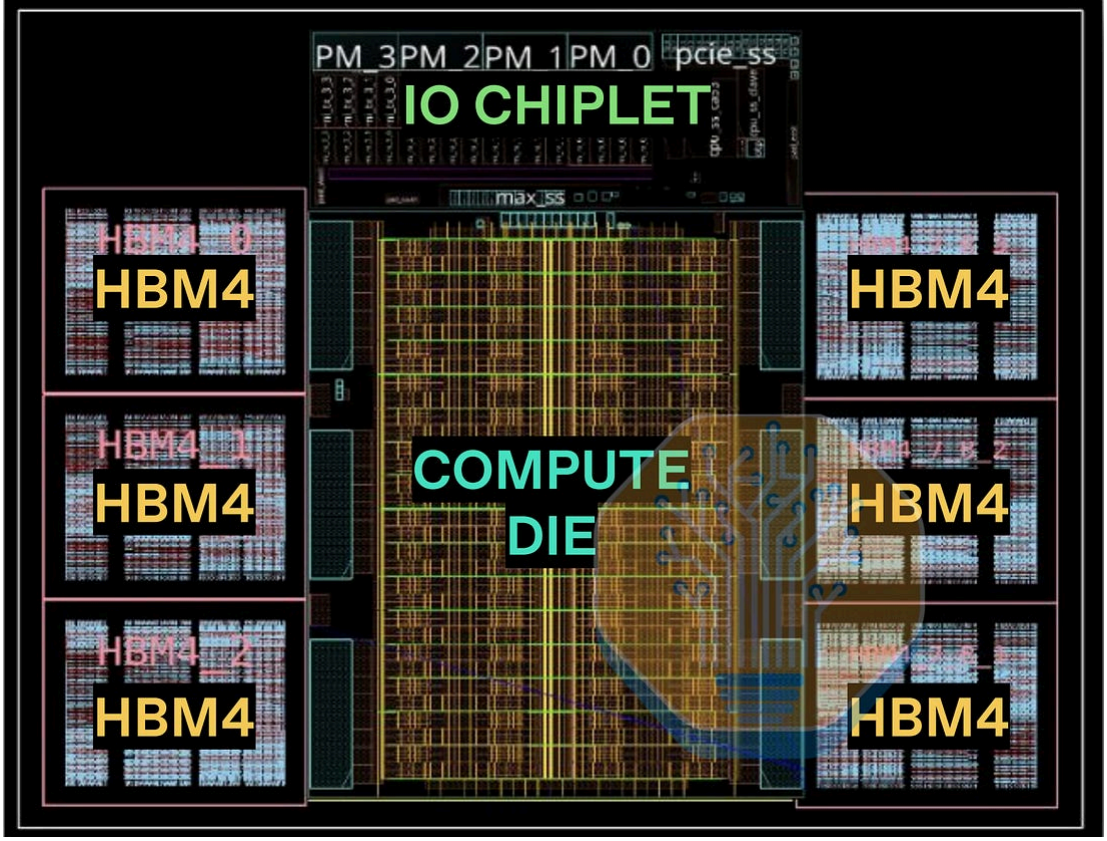
仅从规格来看，它立即成为有力竞争者：

OpenAI Jalapeno Chip Specs						
	FP16 perf	FP8 perf	FP4 perf	HBM capacity	HBM bandwidth	TDP
H100	0.989 PFLOPS	1.979 PFLOPS	-	80 GB	3.35 TB/s	700W
H200	0.989 PFLOPS	1.979 PFLOPS	-	141 GB	4.80 TB/s	700W
MI355X	2.3 PFLOPS	4.6 PFLOPS	9.2 PFLOPS	288 GB	7.987 TB/s	1400W
GB200	2.5 PFLOPS	5 PFLOPS	10 PFLOPS	192 GB	8 TB/s	1200W
GB300	2.5 PFLOPS	5 PFLOPS	15 PFLOPS	288 GB	8 TB/s	1400W
Jalapeno	-	3.4 PFLOPS	13.4 PFLOPS	216 GiB	15.4 TB/s	700W
Rubin	4 PFLOPS	17.5 PFLOPS	35 PFLOPS	288 GB	20 TB/s	1800-2300W
MI450X	10 PFLOPS	20 PFLOPS	40 PFLOPS	432 GB	23.347 TB/s	2500W

Source: SemiAnalysis SemiAnalysis

And the use of HBM4 makes it stand out as comparable to flagship GPUs from NVIDIA and AMD:

而采用 HBM4 使其脱颖而出，可与 NVIDIA 和 AMD 的旗舰 GPU 相媲美：



Source: OpenAI OpenAI

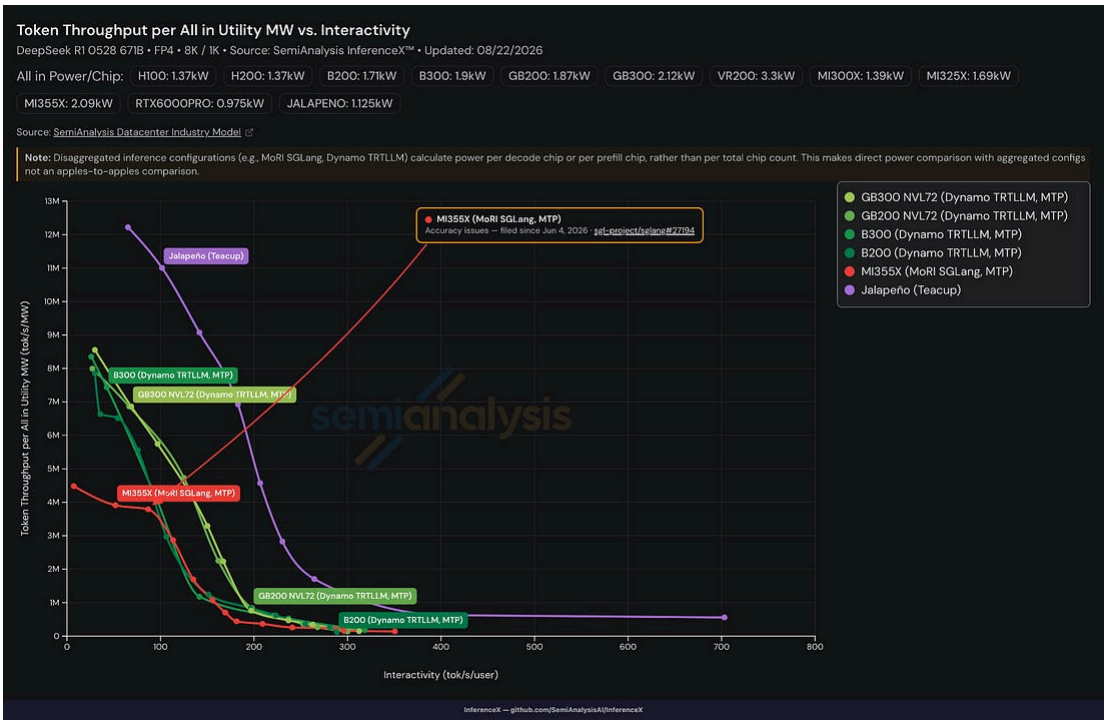
A lot of the media coverage of this chip has followed a few throwaway comments from OpenAI that claim the chip will be optimized for their models in a way that other chips are not. This is wrong. Jalapeño is a generalized inference chip capable of running all sorts of models, and all sorts of workloads, including our benchmark InferenceX, where we ran the benchmark with OpenAI engineers in the lab. As a joke, OpenAI even showed us it running Doom, which was ported to their chip with just Codex prompts.

关于这款芯片的大量媒体报道，都源于 OpenAI 几句随口一提的评论，声称该芯片将针对其模型进行优化，而其他芯片则不具备这种优化。这种说法是错误的。Jalapeño 是一款通用推理芯片，能够运行各类模型和各种工作负载，包括我们的 InferenceX 基准测试——我们曾与 OpenAI 工程师在实验室中共同运行该基准测试。开个玩笑，OpenAI 甚至向我们展示了它在运行《毁灭战士》（Doom），这款游戏仅通过 Codex 提示词就被移植到了他们的芯片上。

The following is our headline perf/W result, looking at token throughput per All-in utility MW. **Jalapeño smokes every other chip**. All this is done without Multi Token Prediction (MTP), while the other chips on the chart are the best performing configs

of each respective SKU, all with MTP.

以下是我们最核心的每瓦性能结果，以每综合公用事业兆瓦（All-in utility MW）的 token 吞吐量衡量。Jalapeño 远超其他所有芯片。上述结果均未使用多 Token 预测（MTP），而图表中其他芯片均为各自 SKU 的最佳配置，且均启用了 MTP。



Source: SemiAnalysis SemiAnalysis

Jalapeño beats Blackwell on perf/W across almost all scenarios without being tuned for any specific point in the curve. It excels not only in low-latency scenarios but also in high-throughput scenarios. A more apples to apples comparison is against Single Token Prediction results, it knocks every competitor out of the water. At low concurrency scenarios, Jalapeño demonstrates remarkable interactivity, hitting over 700 tokens per sec per user at concurrency 1 on the DeepSeek R1 model.

Jalapeño 在几乎所有场景下的每瓦性能（perf/W）都优于 Blackwell，且无需针对曲线中任何特定点进行调优。它不仅擅长低延迟场景，在高吞吐场景下同样表现出色。更公平的对比应是与单 Token 预测（Single Token Prediction）结果的比较，Jalapeño 在这一维度上碾压所有竞争对手。在低并发场景下，Jalapeño 展现出卓越的交互性，在 DeepSeek R1 模型上、并发数为 1 时，每用户每秒可输出超过 700 个 token。

Incredibly, this is all achieved with single-token prediction (STP), no speculative decoding and no prefill-decode disaggregation. In addition to DeepSeek R1, we also got to see some other models, including Kimi-K2.5 and GPT-OSS which ran at approximately 1,400 tok/sec/user. For all models, we confirmed that Jalapeño's

GSM8k evals attained results on par with Nvidia chips.

令人难以置信的是，这一切仅通过单 token 预测（STP）实现，无需投机解码，也无需 prefill-decode 分离。除 DeepSeek R1 外，我们还看到了其他一些模型，包括 Kimi-K2.5 和 GPT-OSS，其运行速度约为 1,400 tok/sec/user。对于所有模型，我们确认 Jalapeño 的 GSM8k 评估结果与 Nvidia 芯片持平。

Some caveats on this. First, all numbers are provided to us by OpenAI. We verified the InferenceX runs in person in the lab, but we did not run the full suite of InferenceX benchmarks nor have we seen AgentX results. AgentX is our preferred suite for comparing chip performance due to the datasets' long context and multi-turn characteristics that reflect the cache behavior of realistic production workflows. Frameworks that perform well on 8k1k may perform worse on AgentX as real production loads stress components like routers, prefix cache mechanisms, cache management, offload infrastructure, etc. These are not tested by single turn 8k1k. Read more about this in out AgentX article.

对此需说明几点。首先，所有数据均由 OpenAI 提供给我们。我们在实验室现场验证了 InferenceX 的运行情况，但并未运行完整的 InferenceX 基准测试套件，也未看到 AgentX 的结果。AgentX 是我们比较芯片性能时优先采用的测试套件，因为其数据集具有长上下文和多轮对话特征，能够反映真实生产工作负载中的缓存行为。在 8k1k 上表现优异的框架在 AgentX 上可能表现更差，因为真实生产负载会对路由器、前缀缓存机制、缓存管理、卸载基础设施等组件产生压力，而这些并非单轮 8k1k 测试所能覆盖。更多详情请参阅我们的 AgentX 文章。

AgentX - InferenceXv3: Does CUDA Moat Hold up in Agentic Inferencing?

AgentX – InferenceXv3：在智能体推理（Agentic Inferencing）中，CUDA 护城河是否依然稳固?

CAM QUILICI, BRYAN SHAN, AND 5 OTHERS

CAM QUILICI、BRYAN SHAN 及其他 5 人 · AUG 24 8 月 24 日

Read full story 阅读全文

Second, we believe that comparison to Blackwell is somewhat incomplete and Jalapeño is really competing against chips like Rubin that also use HBM4. Ver Rubin systems are starting to ship to customers right now, while it will still be

time before OpenAI has anything beyond engineering samples of Jalapeño.

其次，我们认为与 Blackwell 的对比并不完全恰当，也有失公允。Jalapeño 真正对标的是同样采用 HBM4 的 Rubin 等芯片。Vera Rubin 系统目前已经开始向客户出货，而 OpenAI 的 Jalapeño 距离工程样品阶段之外还有相当长的时间。

Thus, performance should really be compared against Rubin, not Blackwell, and in some sense we expect a custom chip like Jalapeño to outperform Blackwell. Vera Rubin NVL72 delivers 5.4x the perf/MW of GB200 NVL72 as we described in our article analyzing the NVIDIA performance claims in their launch with CoreWeave last month. We will compare Jalapeño to Vera Rubin's July performance figures later below.

因此，性能实际上应与 Rubin 而非 Blackwell 进行比较，而且在某种意义上，我们预期像 Jalapeño 这样的定制芯片会优于 Blackwell。Vera Rubin NVL72 的每兆瓦性能（perf/MW）是 GB200 NVL72 的 5.4 倍，正如我们上个月在分析 NVIDIA 与 CoreWeave 发布时性能声明的文章中所描述的那样。我们将在下文将 Jalapeño 与 Vera Rubin 7 月份的性能数据进行对比。

Vera Rubin NVL72 vs GB200 NVL72? Inference TCO & Architecture Analysis

Vera Rubin NVL72 对比 GB200 NVL72？推理总拥有成本（TCO）与架构分析

ALEC IBARRA, BRYAN SHAN, AND 6 OTHERS

ALEC IBARRA、BRYAN SHAN 及其他 6 人 · JUL 23 7 月 23 日

Read full story 阅读全文

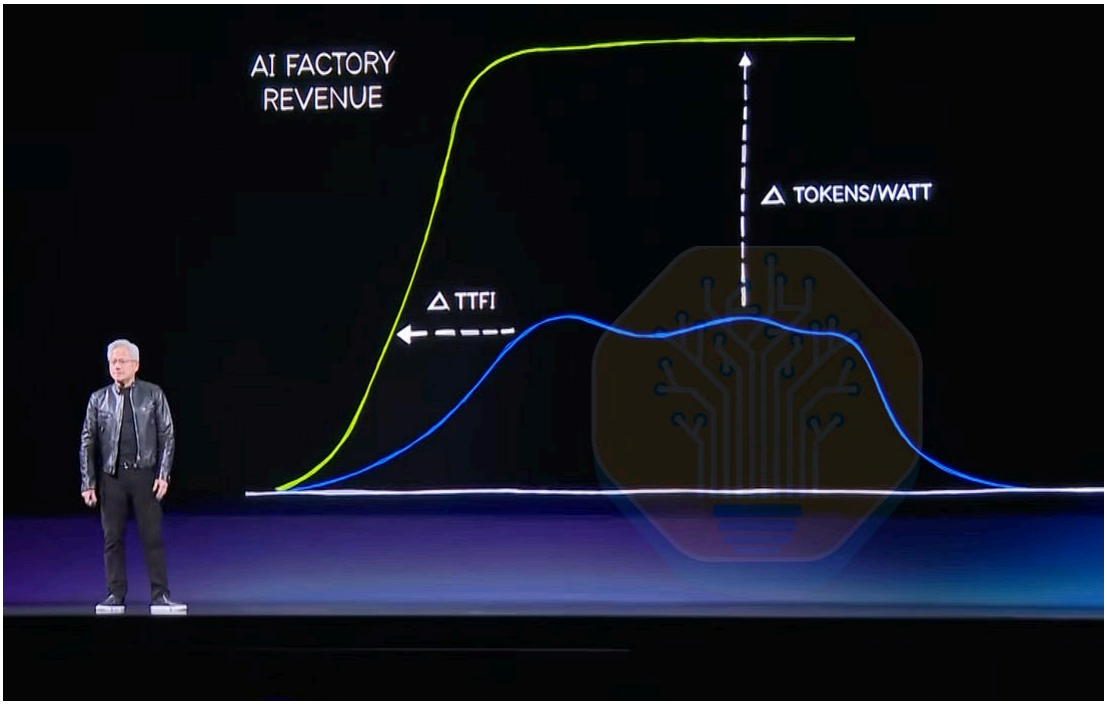
Third, the models being tested are not on the open frontier. NVIDIA and AMD have published results on larger models such as DeepSeek V4 Pro and Kimi K3, using AgentX. The larger the model and the more recent the release, the more complicated it is to bring up on a new chip. With that said the models OpenAI has working on Jalapeno aren't exactly small either.

第三，正在测试的模型并非处于开放前沿。NVIDIA 和 AMD 已使用 AgentX 在 DeepSeek V4 Pro 和 Kimi K3 等更大模型上发布了结果。模型越大、发布越新，在新芯片上启动的复杂度就越高。话虽如此，OpenAI 在 Jalapeno 上运行的模型也并非小模型。

Performance Analysis 性能分析

OpenAI designs for perf/W. The reason is simple: OpenAI is currently limited by datacenter power, not by budget or floorspace, and thus tokens per MW is paramount. At Computex 2026, Jensen said that perf/W, reliability and long lifetime are the core features of future GPUs. To quote: “If you have 1 gigawatt of power, then throughput per watt is revenue”. He also mentioned that choosing the wrong architecture just because the chips are cheaper doesn’t make sense.

OpenAI 以每瓦性能（perf/W）为设计导向。原因很简单：OpenAI 目前受限于数据中心电力，而非预算或机房面积，因此每兆瓦产出的 token 数至关重要。在 Computex 2026 上，黄仁勋表示，每瓦性能、可靠性和长生命周期是未来 GPU 的核心特性。他原话称：“如果你拥有 1 吉瓦电力，那么每瓦吞吐量就是营收。”他还提到，仅仅因为芯片更便宜就选择错误的架构是没有意义的。



Source: Computex 2026 keynote

来源：Computex 2026 主题演讲

This was emphasized by Nvidia during the Vera talk at Hot Chips 2026 while showing the same revenue graph: “The data center is power limited today.” Power matters and drives revenue.

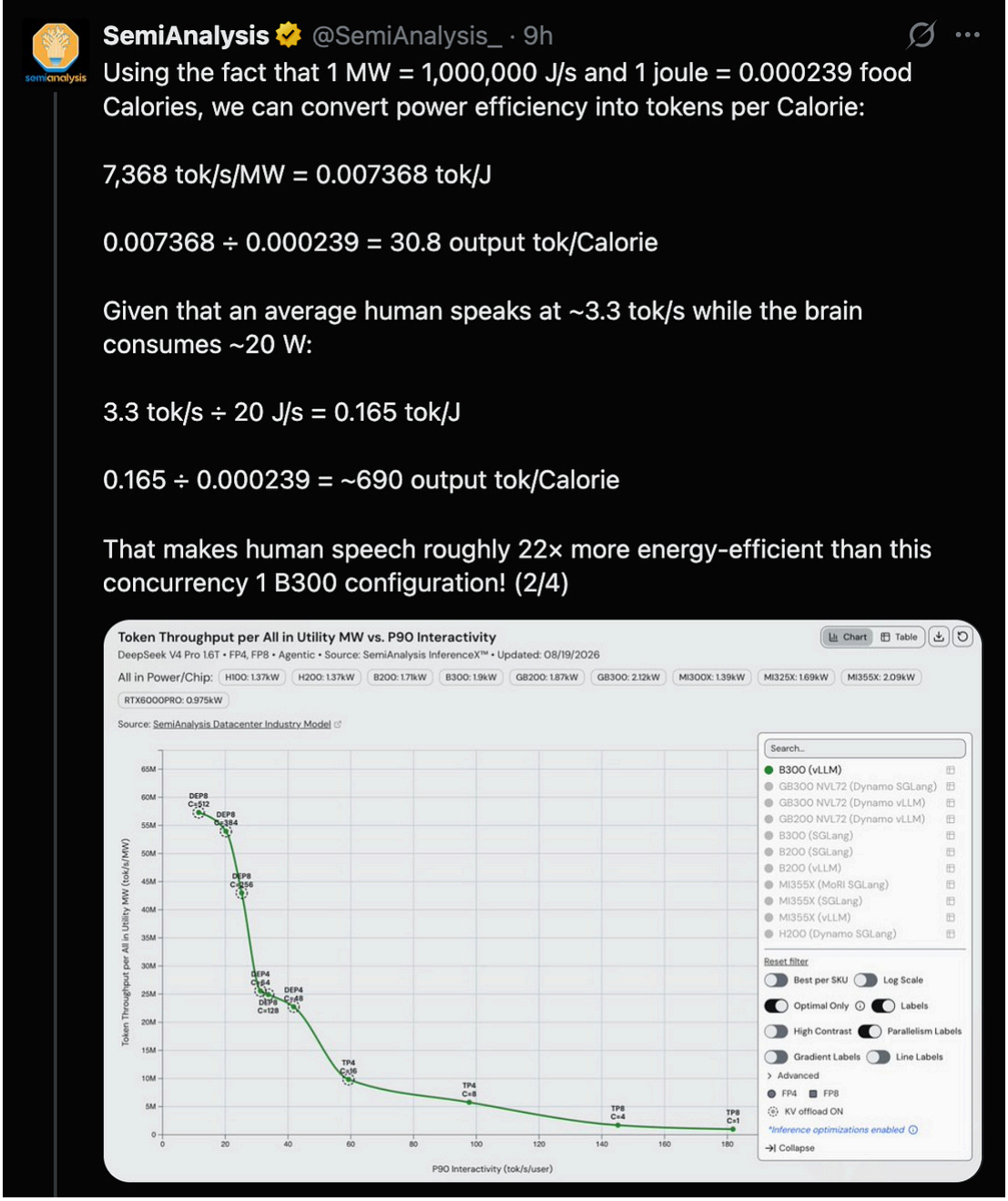
NVIDIA 在 Hot Chips 2026 的 Vera 演讲中展示同一张营收图表时强调：“当前数据中心受限于电力。”电力至关重要，并直接驱动营收。

Operators cannot simply obtain more MW because adding GPUs and adding grid capacity happen on very different timescales. Datacenter power envelopes have constraints such as their utility interconnection, infrastructure, cooling capacity, and UPS/backup-generation design. Grid delays repeatedly outpace hardware and construction timelines, driving the need for BtM (behind-the-meter) power capacity: gas turbines and on-site generators built and located at the data center itself. This capacity sits behind the utility's meter rather than being drawn from the public grid. It lets an operator power a facility without waiting on grid interconnection and utility upgrades, which is exactly why xAI's Colossus 2 relies so heavily on BtM while its actual grid connection lags far behind. Find out more in our Energy model.

运营商无法简单地获取更多 MW，因为增加 GPU 与增加电网容量发生在截然不同的时间尺度上。数据中心电力容量受制于多种约束，包括其电网接入（utility interconnection）、基础设施、冷却能力以及 UPS/备用发电设计。电网延迟一再超过硬件与建设工期，从而催生了对 BtM（表后）电力容量的需求：即在数据中心现场建造并部署的燃气轮机和自备发电机组。该容量位于公用事业电表之后，而非取自公共电网。它使运营商无需等待电网接入和公用事业升级即可为设施供电，这正是 xAI 的 Colossus 2 如此重度依赖 BtM、而其实际电网接入却远远滞后的原因。更多详情请参阅我们的 Energy 模型。

As we wrote in an X post, tok/s/MW reduces to tokens per joule since a watt is a joule per second. This makes tok/s/MW representative of a system's efficiency and ability to convert energy into tokens.

正如我们在 X 帖子中所写，tok/s/MW 可简化为每焦耳 token 数，因为 1 瓦特等于 1 焦耳/秒。这使得 tok/s/MW 能够代表系统将能量转化为 token 的效率与能力。

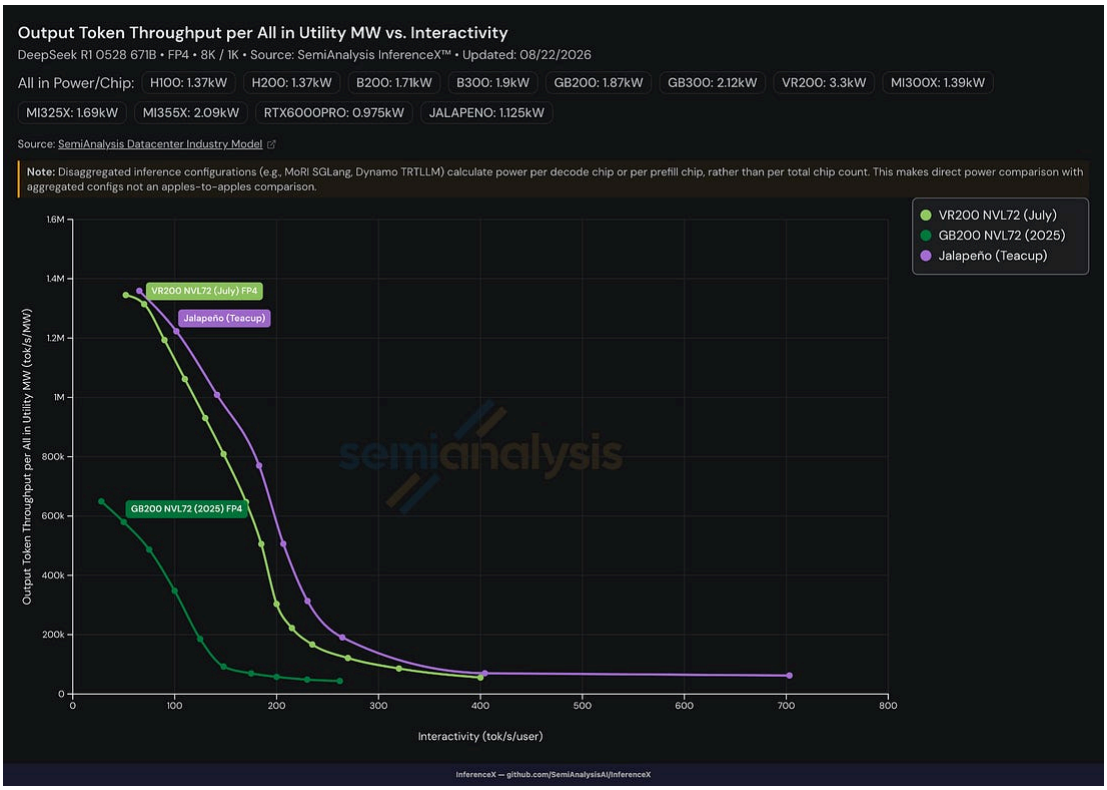


Source: SemiAnalysis SemiAnalysis

On this front, even when compared with Rubin, Jalapeño wins. OpenAI's Jalapeño has STP output token throughput per MW surpassing Vera Rubin's MTP results that NVIDIA and CoreWeave published in July. It also far exceeds GB200's 2025 MTP results. As mentioned in our Vera Rubin article, VR was compared to 2025 GB200 results because that was a similar stage of early bring-up, and comparing to GB200 in 2025 holds software maturity constant. Following this logic, we compare Vera Rubin's latest July 2026 results, GB200 2025 results, and today's Jalapeño results. This is a very valid comparison as these are the best public Rubin numbers, and OpenAI taped out their chip after Rubin. Both OpenAI and Rubin are still immature

thus performance will continue to rise.

在这一方面，即便与 Rubin 相比，Jalapeño 也胜出。OpenAI 的 Jalapeño 每 MW 的 STP 输出 token 吞吐量，超过了 NVIDIA 和 CoreWeave 在 7 月公布的 Vera Rubin 的 MTP 结果，也远超 GB200 在 2025 年的 MTP 表现。正如我们在 Vera Rubin 文章中所提到的，VR 当时是与 2025 年的 GB200 结果进行比较，因为两者处于相似的早期 bring-up 阶段，且与 2025 年的 GB200 对比可保持软件成熟度不变。遵循这一逻辑，我们将 Vera Rubin 最新的 2026 年 7 月结果、GB200 的 2025 年结果以及今日的 Jalapeño 结果放在一起比较。这是一组非常有效的对比，因为这些是公开可得的 Rubin 最佳数据，且 OpenAI 的芯片是在 Rubin 之后完成 tape-out 的。OpenAI 与 Rubin 目前均尚未成熟，因此性能还将继续提升。



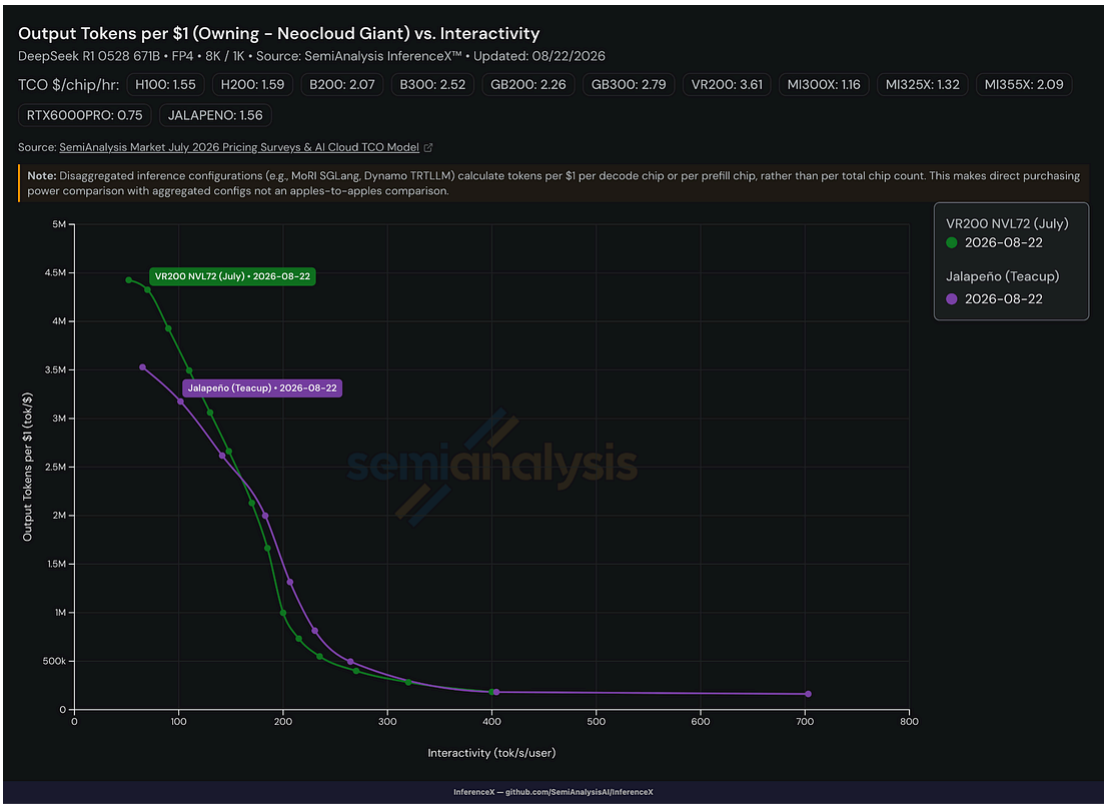
Source: OpenAI, SemiAnalysis

来源：OpenAI、SemiAnalysis

On perf/TCO, Vera Rubin and Jalapeño are head-to-head, producing almost the same number of output tokens per $. However, as previously mentioned, **Jalapeño's results are obtained without speculative decoding** and Vera Rubin's results use speculative decoding. Speculative decoding adds a 3x+ improvement is cost per token. When speculative decoding is implemented on Jalapeño, this will mean Jalapeño will serve much lower cost tokens. Of course, part of this TCO advantage comes from trading Nvidia's high margins for Broadcom's lower (though still high)

margins. But this is not all of it. For example, Meta and Microsoft's AI ASIC programs not getting off the ground despite being at it for much longer shows that cost is only one part of the equation. For Jalapeño's full TCO breakdown, see the SemiAnalysis AI Cloud TCO model.

在性能/TCO（总拥有成本）方面，Vera Rubin 与 Jalapeño 势均力敌，每美元产出的输出 token 数几乎相同。然而，如前所述，Jalapeño 的结果是在未使用投机解码（speculative decoding）的情况下取得的，而 Vera Rubin 的结果则使用了投机解码。投机解码在每 token 成本上带来了 3 倍以上的提升。当 Jalapeño 实现投机解码后，这意味着 Jalapeño 将以远更低的成本提供 token。当然，这一 TCO 优势部分源于以 NVIDIA 的高利润率换取 Broadcom 较低（但仍属高位）的利润率。但这并非全部原因。例如，Meta 和 Microsoft 的 AI ASIC 项目尽管投入时间更长却未能落地，这表明成本只是方程式中的一部分。关于 Jalapeño 的完整 TCO 拆分，请参阅 SemiAnalysis AI Cloud TCO 模型。



Source: OpenAI, SemiAnalysis

来源：OpenAI、SemiAnalysis

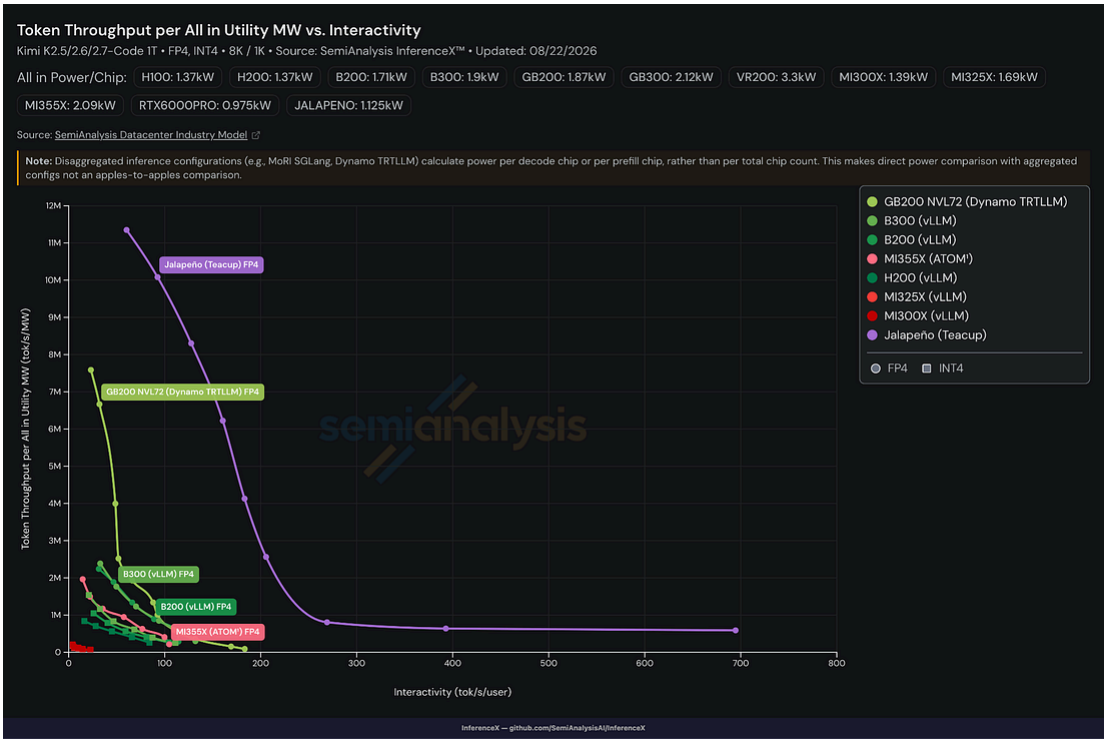
Architecturally, OpenAI chose not to disaggregate prefill and decode (PD) across separate chip pools. The draft model and main model share the same chips and fabric, a design philosophy that trades some theoretical efficiency for practical operations. The motivation is that the workload mix changes over time, for example the ratio of input to cache write to cache read to output tokens has changed significantly as we

have moved through the three eras of models (knowledge, reasoning, and agentic, as discussed in our recent article). Therefore, picking a fixed amount of heterogenous prefill silicon and decode silicon up front can lead to inefficiencies over time. OpenAI chooses a homogenous pool in this architecture and tries to make the chip perform well on everything.

架构上，OpenAI 选择不将 prefill 和 decode（PD）分离到不同的芯片池中。草稿模型和主模型共享相同的芯片和互联结构，这一设计理念以部分理论效率换取实际运营的便利。其动机在于工作负载组合会随时间变化，例如，随着我们经历模型发展的三个时代（知识、推理和智能体，详见我们近期文章），输入 token、缓存写入、缓存读取与输出 token 的比例已发生显著变化。因此，预先固定分配异构的 prefill 和 decode 芯片数量，长期来看可能导致效率低下。OpenAI 在此架构中选择同质化芯片池，并力求芯片在各种任务上均表现优异。

And it does. On Kimi K2.5 (which Cursor Composer 2.5 is based on), Jalapeño reaches nearly 700tok/s/user and more than 9x the next best performing chip at 100tok/s/user.

事实也确实如此。在 Kimi K2.5（Cursor Composer 2.5 基于此模型）上，Jalapeño 在每用户 100tok/s 的负载下，吞吐量接近 700tok/s/用户，是性能次优芯片的 9 倍以上。

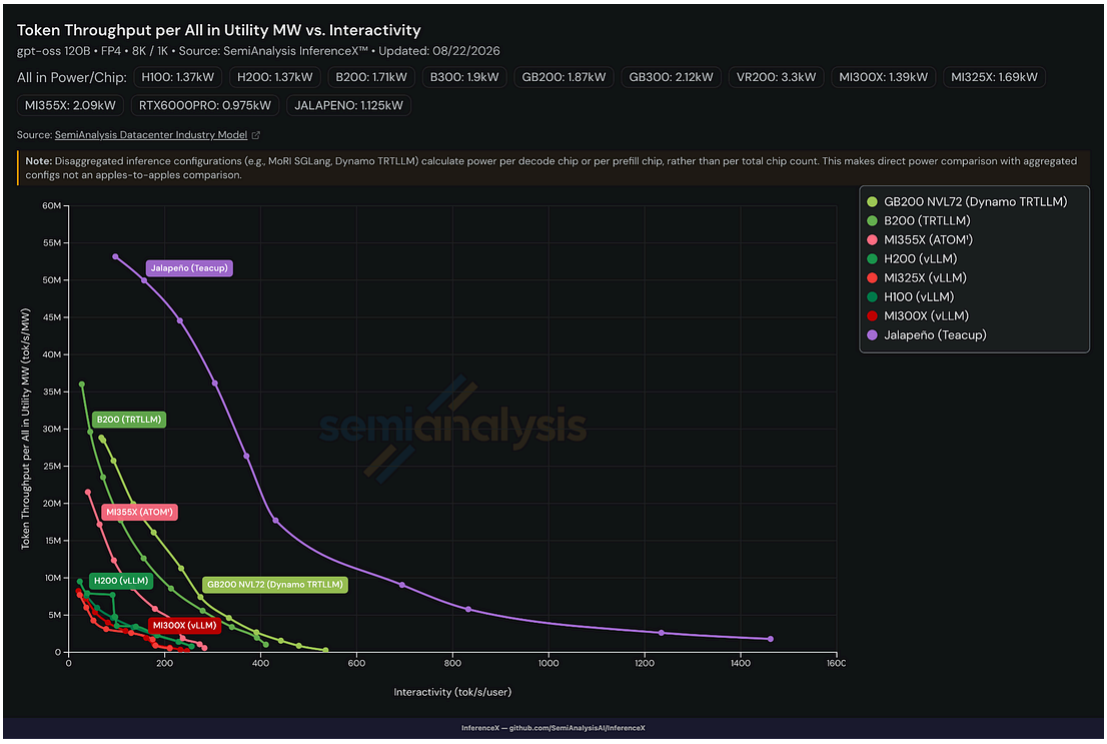


Source: OpenAI, SemiAnalysis

来源：OpenAI、SemiAnalysis

On GPT-OSS, it's another bloodbath. Jalapeño's iso-interactivity throughput per MW is nearly double GB200's highest throughput point and more than 50x GB200's concurrency 1 point. The higher concurrency Jalapeño points use EP8.

在 GPT-OSS 上，又是一场血洗。Jalapeño 每 MW 的 iso-interactivity 吞吐量几乎是 GB200 最高吞吐量点的两倍，且超过 GB200 并发 1 点的 50 倍以上。Jalapeño 更高并发点采用 EP8。



Source: OpenAI, SemiAnalysis

来源：OpenAI、SemiAnalysis

These results are impressive! However, we have to nitpick: they're just 8k1k, a much easier workload to tune for, and there are no AgentX runs yet. As mentioned in our AgentX article, multiturn, long context workloads stress much more aspects of the serving stack, such as routers and prefix cache. Many more optimizations are needed to excel in agentic workloads. Read more about this in the AgentX article.

这些结果令人印象深刻！不过，我们还是要吹毛求疵一下：这些只是 8k1k（8K 上下文、单序列）的测试，是更容易调优的工作负载，而且目前还没有 AgentX 的跑分结果。正如我们在 AgentX 文章中所提到的，多轮对话、长上下文工作负载会对服务栈的更多环节产生压力，例如路由器和前缀缓存（prefix cache）。要在智能体工作负载中表现出色，还需要更多的优化。更多详情请参阅 AgentX 文章。

AgentX - InferenceXv3: Does CUDA Moat Hold up in Agentic Inferencing?

AgentX – InferenceXv3：在智能体推理（Agentic Inferencing）中，CUDA 护城河是否依然稳固?

CAM QUILICI, BRYAN SHAN, AND 5 OTHERS

CAM QUILICI、BRYAN SHAN 及其他 5 人

· AUG 24 8 月 24 日

Read full story 阅读全文

Digging into the specs and architecture

深入解析规格与架构

All these results were gathered on the A0 stepping of Jalapeño, just 9 months into the program. But there is already a B0 stepping that is currently in the fab! B0 has optimizations that deliver roughly a 25% perf-per-watt improvement over the earlier A0 silicon. Specifically, the B0 stepping delivers 13.4 PFLOPs of MXFP4 on a single reticle-sized compute die that is manufactured on TSMC's N3P. This compares to 17.5 PFLOPs of dense Rubin NVFP4 for a single Rubin compute die that is similar size and on the same node.

以上所有结果均基于 Jalapeño 的 A0 stepping 测得，此时距项目启动仅 9 个月。但目前已有 B0 stepping 正在晶圆厂中流片！B0 包含优化设计，相较早期 A0 硅片可实现约 25%的每瓦性能提升。具体而言，B0 stepping 在单个与 Rubin 计算 die 尺寸相当、同样采用台积电 N3P 工艺的 reticle 级计算 die 上，可实现 13.4 PFLOPs 的 MXFP4 算力。相比之下，单个 Rubin 计算 die 在相同尺寸和节点下，其密集 NVFP4 算力为 17.5 PFLOPs。

This is more respectable considering Jalapeño's TDP is only 700W compared to Rubin's at 900-1,150W per compute die. As Jalapeño is geared towards inference rather than training, it is understandable that OpenAI doesn't need to push TDPs higher to maximize FLOPs, but regardless the above shows that Jalapeño delivers

respectable peak theoretical FLOPs.

考虑到 Jalapeño 的 TDP 仅为 700W，而 Rubin 每个计算 die 的 TDP 为 900–1,150W，这一结果更显体面。由于 Jalapeño 面向推理（inference）而非训练（training），OpenAI 无需将 TDP 推高以最大化 FLOPs，这可以理解；但无论如何，上述数据表明 Jalapeño 在理论峰值 FLOPs 上表现可观。

When compared directly to other accelerators, Jalapeño has the highest HBM bandwidth per watt, and the highest FLOPs per watt, comparable to the 1,800W Rubin Max-Q configuration:

与其他加速器直接对比时，Jalapeño 拥有最高的每瓦 HBM 带宽和每瓦 FLOPs，与 1,800W 的 Rubin Max-Q 配置相当

OpenAI Jalapeno Chip Specs									
	FP16 perf	FP8 perf	FP4 perf	HBM capacity	HBM bandwidth	TDP	HBM bw per W	FLOPs per W	perf:bw ratio
H100	0.989 PFLOPS	1.979 PFLOPS	-	80 GB	3.35 TB/s	700W	4.79	2.8	591
H200	0.989 PFLOPS	1.979 PFLOPS	-	141 GB	4.80 TB/s	700W	6.86	2.8	412
MI355X	2.3 PFLOPS	4.6 PFLOPS	9.2 PFLOPS	288 GB	7.987 TB/s	1400W	5.71	6.6	1152
GB200	2.5 PFLOPS	5 PFLOPS	10 PFLOPS	192 GB	8 TB/s	1200W	6.67	8.3	1250
GB300	2.5 PFLOPS	5 PFLOPS	15 PFLOPS	288 GB	8 TB/s	1400W	5.71	10.7	1875
Jalapeno	-	3.4 PFLOPS	13.4 PFLOPS	216 GiB	15.4 TB/s	700W	22	19.1	870
Rubin	4 PFLOPS	17.5 PFLOPS	35 PFLOPS	288 GB	20 TB/s	1800-2300W	11.1 - 8.7	19.4 - 15.2	1750
MI450X	10 PFLOPS	20 PFLOPS	40 PFLOPS	432 GB	23.347 TB/s	2500W	7.86	16	2034

Source: SemiAnalysis SemiAnalysis

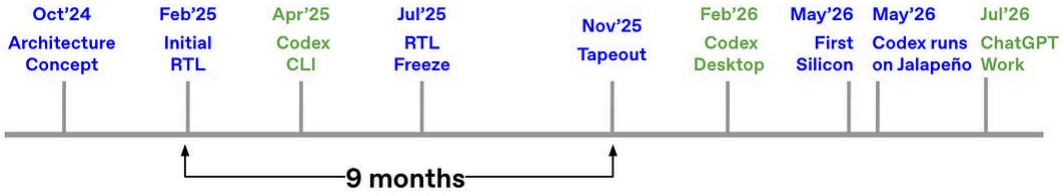
Off-package I/O is provided by an N3E I/O chiplet with 32 lanes of 800G SerDes, for the compute fabric, with 24 lanes (600GB/s) being used for local scale-up within the rack, and 8 lanes (200GB/s) for global scale-up which is the 2,048 XPU multi-rack domain. PCIe Gen 5 is used for system I/O to connect to the x86 host CPU.

片外 I/O 由 N3E I/O chiplet 提供，配备 32 通道 800G SerDes，用于计算 fabric；其中 24 通道（600GB/s）用于机架内本地 scale-up，8 通道（200GB/s）用于全局 scale-up，即 2,048 XPU 多机架域。系统 I/O 采用 PCIe Gen 5 连接 x86 主机 CPU。

Jalapeño will ship with HBM4, making this chip one of the relatively early adopters after Nvidia and AMD, even beating the established TPU and Trainium programs. As one of the key architectural principles behind Jalapeño is getting the most out of HBM bandwidth, settling for anything but the best HBM would run counter to that goal. This results in 15.4TB/s of memory bandwidth per package which bests all the other accelerators shipping that are using HBM3E. The 15.4TB/s bandwidth shows its HBM4 can hit 10Gbps pin speeds, which would give it a slight edge over the 9.6Gbps Nvidia is getting out of its HBM4 in Rubin. The HBM is likely provided by

Samsung.

Jalapeño 将搭载 HBM4 出货，使其成为继 NVIDIA 和 AMD 之后相对较早采用 HBM4 的芯片，甚至领先于成熟的 TPU 和 Trainium 项目。由于 Jalapeño 的关键架构原则之一是最大化利用 HBM 带宽，因此若退而求其次选择非最优的 HBM，将与这一目标背道而驰。这使得每个封装可提供 15.4TB/s 的内存带宽，超越了所有其他采用 HBM3E 的在售加速器。15.4TB/s 的带宽表明其 HBM4 可实现 10Gbps 的引脚速率，略高于 NVIDIA 在 Rubin 中 HBM4 所达到的 9.6Gbps。该 HBM 很可能由三星提供。



Source: OpenAI OpenAI

OpenAI taped out Jalapeño in November 2025, or more specifically, this was a tape out of the CoWoS design, not just the top die silicon. Within 9 months of that Nov 2025 tapeout, and with only 3 months of bring-up on actual silicon, OpenAI has already delivered very good results with Jalapeño. This is all the more impressive as the team is starting from zero on the software stack.

OpenAI 于 2025 年 11 月完成了 Jalapeño 的 tape out，更准确地说，这是 CoWoS 封装的 tape out，而不仅仅是顶层芯片（top die）的硅片。自 2025 年 11 月那次 tape out 以来，在短短 9 个月内，且实际芯片 bring-up 仅用了 3 个月，OpenAI 已经在 Jalapeño 上取得了非常好的成果。考虑到该团队在软件栈方面是从零起步，这一成绩尤为令人印象深刻。

Meanwhile, Rubin's CoWoS tape out was completed in October 2025, a month earlier, and yet the only early results we have seen are from CoreWeave's engineering samples. Nvidia has not let us test and release benchmarks in the same way that OpenAI has, indicating their chip software is still immature. The CUDA moat is

potentially dead given how fast OpenAI can bring up new models on their silicon.

与此同时，Rubin 的 CoWoS 流片已于 2025 年 10 月完成，比预期提前一个月，但迄今为止我们看到的早期结果仅来自 CoreWeave 的工程样品。NVIDIA 并未像 OpenAI 那样允许我们测试并发布基准测试结果，这表明其芯片软件栈仍不成熟。鉴于 OpenAI 能如此迅速地在自研芯片上完成新模型的部署，CUDA 的护城河可能已经名存实亡。

They are still far from optimized and we can see that generally Jalapeño has delivered better numbers. We don't think that Nvidia hardware is inferior, but more so that Jalapeño's software bring-up has progressed more quickly than Nvidia's. This speaks to the power of hardware/software co-design, which is the main area where a cracked frontier lab ASIC team can excel over more established merchant silicon players. Counterintuitively, starting from scratch may also have benefited OpenAI as it could make clean-sheet architectural decisions without worrying about backwards compatibility or older software versions.

它们距离优化完成仍有很大差距，总体来看，Jalapeño 交出的数据更胜一筹。我们不认为 NVIDIA 的硬件逊色，而更倾向于认为 Jalapeño 的软件栈推进速度比 NVIDIA 更快。这印证了软硬件协同设计（hardware/software co-design）的价值，而这正是顶尖前沿实验室的 ASIC 团队相较于成熟商用芯片厂商最可能实现超越的核心领域。反直觉的是，从零起步或许也让 OpenAI 受益，因为它可以做出白纸式的架构决策，无需顾虑向后兼容性或旧版软件。

While OpenAI has engineering samples of Jalapeño, production is currently scheduled to gradually ramp over 2027 with most of the output currently scheduled for the end of next year. For more details of unit volumes and ASPs, see the SemiAnalysis Accelerator Model.

虽然 OpenAI 已拥有 Jalapeño 的工程样品，但量产目前计划在 2027 年内逐步爬坡，大部分产出目前定于明年年底交付。有关出货量与 ASP 的更多细节，请参阅 SemiAnalysis Accelerator Model。

Suffice to say, OpenAI Jalapeno is a real high volume ASIC.

可以说，OpenAI Jalapeno 是一款真正的高吞吐量 ASIC。

When compared against Rubin's timeline, Jalapeño's is shockingly quick. As shown earlier, Jalapeño's results beat Rubin's despite Rubin's head start.

与 Rubin 的时间线相比，Jalapeño 的推进速度快得惊人。如前所述，尽管 Rubin 起步更早，但 Jalapeño 的结果仍优于 Rubin。

Source: SemiAnalysis SemiAnalysis

Jalapeno Architecture Jalapeno 架构

Digging into the architecture now, the chip's matrix engine uses MXFP numerical formats and a weight stationary systolic array, similar to TPU. But when compared directly to TPU, it has support for smaller shapes / dimensions, meaning that it doesn't have weird performance cliffs that get exposed by awkwardly shaped matmuls on bigger systolics.

现在深入分析架构层面，该芯片的矩阵引擎采用 MXFP 数值格式和权重驻留脉动阵列（weight stationary systolic array），与 TPU 类似。但直接对比 TPU 时，它支持更小的形状/维度，这意味着它不会像大型脉动阵列那样，因形状不规整的矩阵乘法（matmul）而暴露出奇怪的性能悬崖。

It also has 64-bit scalar cores and FP32/INT32 vector cores. OpenAI has also invested in redundancy at the tray level and has yield harvesting built in at the core and channel level. They claim that AI assistance in chip design delivered an 8% reduction in SIMD area and a 10% reduction in matrix-engine area during design. While they did not clarify the exact process/voltage/temperature (PVT) conditions, they also mentioned the AI-assisted blocks improved timing and power over the initial blocks.

它还配备了 64 位标量核心和 FP32/INT32 向量核心。OpenAI 还在托盘层面投资了冗余设计，并在核心和通道层面内置了良率收割机制。他们声称，在芯片设计过程中，AI 辅助设计使 SIMD 面积减少了 8%，矩阵引擎面积减少了 10%。虽然他们没有明确具体的工艺/电压/温度（PVT）条件，但他们也提到，AI 辅助模块在时序和功耗方面优于初始模块。

The Jalapeño architecture design focuses on eliminating memory movement of KVCache and weights as well as fixed latencies and overheads in order to make it

possible to get closer to the raw peak flops/bandwidth even for small batches or shapes as compared to other accelerators.

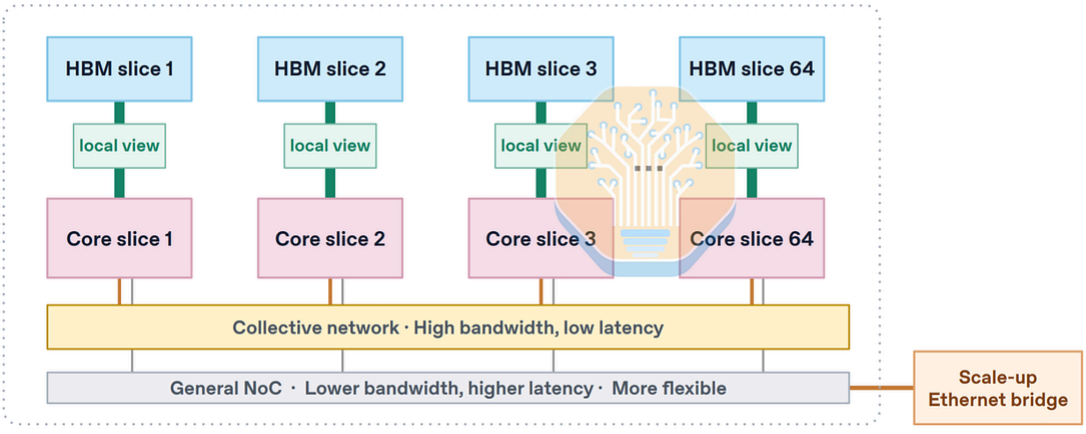
Jalapeño 架构的设计重点在于消除 KVCache 与权重的内存搬运，以及固定的延迟和开销，从而即便面对小批次（small batches）或不规则形状（shapes），也能比其他加速器更接近原始峰值算力/带宽（raw peak flops/bandwidth）。

The cores and the HBM are divided into slices, where each core slice has a low-latency local view on its own slice of HBM. Synchronization between slices occurs on a high-bandwidth dedicated collective network. This minimal memory hierarchy already gives Jalapeño a big potential advantage over GPUs, where memory accesses must traverse a complicated memory system, resulting in large latencies that must be amortized or hidden over larger shapes.

核心（cores）与 HBM 被划分为多个切片（slices），每个核心切片对其自身的 HBM 切片拥有低延迟的本地视图。切片之间的同步通过高带宽的专用集合通信网络（collective network）完成。这一极简内存层级已使 Jalapeño 相比 GPU 具备巨大潜在优势——在 GPU 上，内存访问必须穿越复杂的内存系统，导致较大延迟，且这些延迟必须通过更大的形状来摊还（amortized）或隐藏。

This choice is feasible because with careful placement of weights and KVs, synchronization between cores can be restricted to limited, known high-bandwidth comms such as tensor-parallel communication that can be overlapped with compute.

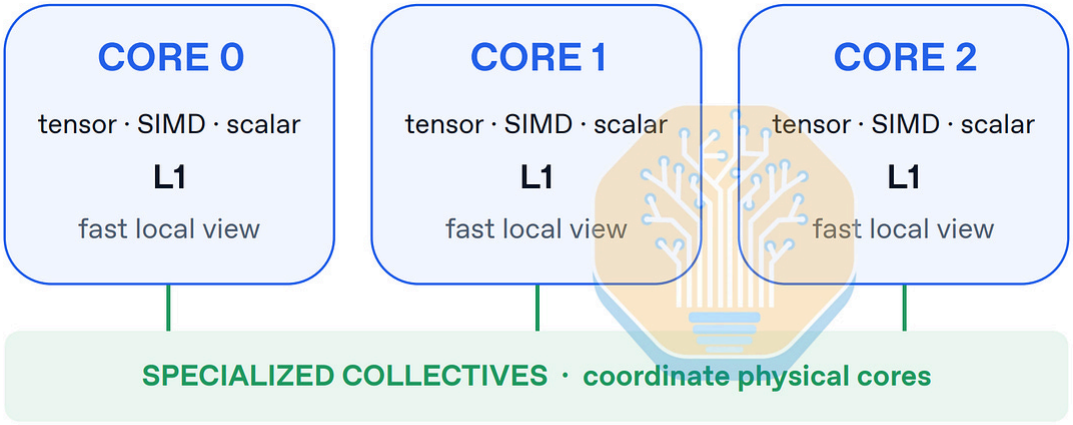
这一选择之所以可行，是因为通过对权重和 KV（KVs）进行精心放置，核心之间的同步可以被限制在有限且已知的高带宽通信范围内，例如可与计算重叠（overlapped with compute）的张量并行（tensor-parallel）通信。



Source: OpenAI OpenAI

There is also an additional general NoC which is used for general comms and to access the scale-up network. In general OpenAI saves huge power and gets big performance gains with a simplified NOC and memory subsystem vs Nvidia and Google.

此外，还有一个用于通用通信及访问 scale-up 网络的通用 NOC。总体而言，与 NVIDIA 和 Google 相比，OpenAI 通过简化的 NOC 和内存子系统，在功耗和性能上获得了巨大优势。



Source: OpenAI OpenAI

At the core level, OpenAI describes an out-of-order (OoO) core with an L1 cache. This is a large divergence from the pattern we have seen in other accelerators, all of which instead use software-managed scratchpad commonly paired with some async DMA support. Again, the argument being made here is that this allows Jalapeño to avoid fixed overheads such as barrier latencies, which on other accelerators (such as GPUs) need to be hidden or amortized over with higher work per core, and make it harder to get close to the raw peak bandwidth/flops.

在核心层面，OpenAI 描述了一个带有 L1 缓存的无序执行（out-of-order, OoO）核心。这与我们在其他加速器中看到的模式存在重大差异——后者均采用软件管理的 scratchpad（便笺式存储器），并通常搭配异步 DMA 支持。这里的论点再次是：这种设计使 Jalapeño 能够避免固定的开销，例如 barrier 延迟；而在其他加速器（如 GPU）上，这些开销需要通过每个核心更高的工作量来隐藏或摊销，从而难以接近原始峰值带宽/算力（FLOPs）。

The tradeoff is that Jalapeño therefore relies on good prefetching to ensure timely arrivals of memory requests, which is less predictable and more difficult to reason about. However, with Codex in a good harness with access to detailed tracing, it is

likely that finding the optimal kernel with the best prefetching for a given shape requires little human intervention. We think that is exactly what OpenAI has done to bring up DeepSeek R1, Kimi K2.5, and GPT-OSS so quickly.

其代价在于，Jalapeño 因此依赖良好的预取（prefetching）来确保内存请求及时到达，而这 predictability 较低，也更难进行推理分析。然而，在 Codex 配合良好 harness 并具备详细 tracing 能力的情况下，针对特定 shape 找到具备最优预取策略的 kernel，很可能只需极少的人工干预。我们认为，OpenAI 正是借此在如此短的时间内推出了 DeepSeek R1、Kimi K2.5 和 GPT-OSS。

The cores also have support for "small" matrix dimensions, which (depending on how small) should make it more general across different model and batch dimensions less sensitive to matrix dimension alignment, padding overhead, and tiling inefficiency. For instance, TPUs, Trainium, and Etched chips have very large systolic arrays which can require large batches or exactly-divisible model dimensions to avoid tiling inefficiencies.

这些核心还支持“小”矩阵维度，这（取决于具体有多小）应使其在不同模型和批量维度上更具通用性，对矩阵维度对齐、填充开销和分块（tiling）低效的敏感度更低。例如，TPU、Trainium 和 Etched 芯片拥有非常大的脉动阵列（systolic array），这可能需要大批量或能被精确整除的模型维度，以避免分块低效问题。

With Jalapeño, OpenAI has focused on eliminating fixed latencies in the system to allow for as-close-to-roofline performance as possible across all areas of the pareto curve. In theory, this could give them advantages over the GPU at multiple operating points:

借助 Jalapeño，OpenAI 专注于消除系统中的固定延迟，以便在帕累托曲线（pareto curve）的所有区域尽可能实现接近性能上限的表现。理论上，这可以让他们在多个运行点上相对于 GPU 获得优势：

- Much better upper-bound performance on low-latency/small-batch inference, which on GPUs is limited by many fixed overheads such as launch latencies, barrier latencies, memory system latency

  在低延迟/小批量推理场景下实现更优的上限性能，而在 GPU 上，此类场景受限于诸多固定开销，如启动延迟、同步屏障延迟、内存系统延迟

- Some potential to achieve closer to the hardware roofline even for large-batch or long-context

  即便在大批量或长上下文场景下，仍有可能更接近硬件性能上限（roofline）

This comes with the caveat that even if the upper-bound performance is available in theory, it may be more difficult to realize that performance for real kernels. So it seems the approach is:

需要指出的是，即便理论上存在上限性能，对于实际内核而言，实现该性能可能更为困难。因此，其思路似乎是：

1. Design for the highest upper-bound performance across all workload shapes

   针对所有工作负载形态设计最高的上限性能

2. Let Codex do the tedious work of finding the kernels that achieve that upper bound

   让 Codex 完成寻找能达到该上限的内核的繁琐工作

Judging by the extremely fast turnaround for the OpenAI team to bring up InferenceX workloads on Jalapeño, we are optimistic about this approach.

从 OpenAI 团队在 Jalapeño 上快速部署 InferenceX 工作负载的极短周期来看，我们对这一路径持乐观态度。

If Jalapeño is a success, it will be a strong signal that the industry's obsession over programming models and perfect, universal compilers are invalidated by frontier AI models.

若 Jalapeño 取得成功，这将是一个强烈信号，表明业界对编程模型及完美通用编译器的执着追求，已被前沿 AI 模型所颠覆。

Software 软件

OpenAI writes Jalapeño kernels like assembly. Each kernel gets hand-tuned code, some running to ~3,000 lines, backed by correctness checks and a custom sanitizer. Early kernel work was human-in-the-loop rather than fully automated, but this shifted with a more scaled-up, internal version of Codex, one which OpenAI plans to pitch to enterprise customers. The internal serving engine is called "Teacup". Interestingly, **OpenAI had no internal implementation of MLA kernels until they**

benchmarked DeepSeek with InferenceX. The ability for Codex to write functional and efficient kernels so quickly (without any of OpenAI's kernel engineering team intervening) shows the software pipeline's developmental ability.

OpenAI 编写 Jalapeño 内核的方式如同编写汇编代码。每个内核都经过手工调优，部分代码长达 1,000 至 3,000 行，并辅以正确性检查及自定义 sanitizer。早期内核开发采用人机协同（human-in-the-loop）而非全自动化，但随着内部版本 Codex 的规模化扩展，这一模式发生了转变——OpenAI 计划将该版本 Codex 推销给企业客户。内部推理引擎名为“Teacup”。有趣的是，在 OpenAI 使用 InferenceX 对 DeepSeek 进行基准测试之前，其内部并未实现 MLA 内核。Codex 能够在没有任何 OpenAI 内核工程团队干预的情况下，如此迅速地编写出功能完备且高效的内核，这充分展示了该软件流水线的开发能力。

OpenAI programs Jalapeño with Gluon. Gluon is OpenAI's kernel programming language. Built on top of Triton, Gluon preserves Triton's SPMD (Single Program Multiple Data) programming model, but it exposes **low-level programming abstractions**. For example, for NVIDIA GPUs, it offers APIs that map to PTX instructions, including MMA instructions, TMA instructions, mbarrier mechanisms, and many more. The most unique abstraction Gluon provides is **the layout**. Generally speaking, a layout defines a mapping between a hardware resource (e.g. 5th register of warp 9) and a tensor element (e.g. tensor element on row 6 column 7). Gluon's layout abstraction is based on Linear Layouts, a type of layout algebra OpenAI invented. Linear Layouts mathematically formalizes what a layout is and provides tools to operate on layouts. This enables many features, such as provably correct layout conversions and optimal memory swizzling.

OpenAI 推出 Jalapeño，基于 Gluon 构建。Gluon 是 OpenAI 的 kernel 编程语言。Gluon 构建在 Triton 之上，保留了 Triton 的 SPMD（单程序多数据）编程模型，但暴露了底层编程抽象。例如，针对 NVIDIA GPU，它提供映射到 PTX 指令的 API，包括 MMA 指令、TMA 指令、mbarrier 机制等。Gluon 提供的最独特抽象是 layout（布局）。一般而言，layout 定义了硬件资源（如 warp 9 的第 5 个寄存器）与张量元素（如第 6 行第 7 列的张量元素）之间的映射关系。Gluon 的 layout 抽象基于 Linear Layouts，这是 OpenAI 发明的一种 layout 代数。Linear Layouts 以数学形式正式定义了 layout 的概念，并提供了操作 layout 的工具。这带来了诸多特性，例如可证明正确的 layout 转换以及最优的内存 swizzling。

In terms of Jalapeño's programming model, each Gluon program maps to a persistent thread. We believe this hints that Jalapeño suits the **persistent kernel programming pattern**, where each program executes on multiple tiles, and the programmer, rather than the hardware scheduler, assigns the work. OpenAI mentioned TensorInfo, an abstraction that explicitly encodes layouts. This is likely the set of layouts designed for Jalapeño, which will be powered by Linear Layouts. Finally, each core offers data prefetching and decoupled out-of-order units. For example, a user might program a wait on a prefetched data, which is locked behind a semaphore.

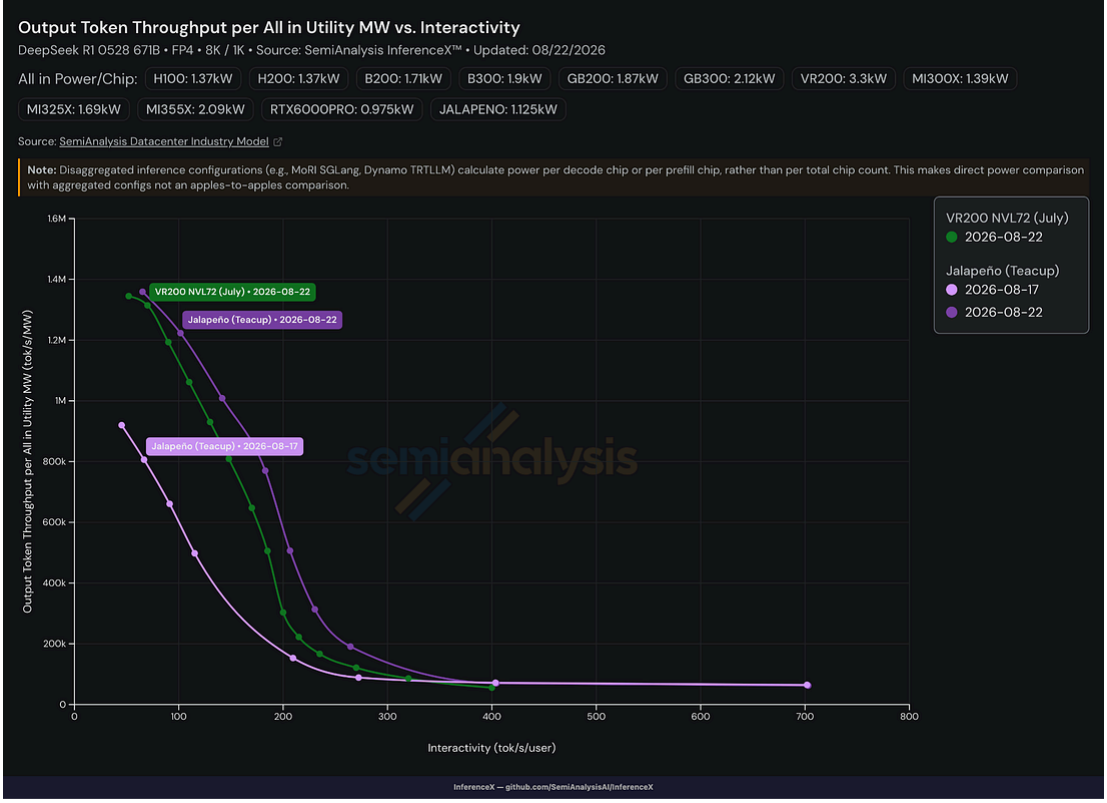
在 Jalapeño 的编程模型方面，每个 Gluon 程序映射到一个持久线程（persistent thread）。我们认为这暗示 Jalapeño 适用于持久内核（persistent kernel）编程模式，即每个程序在多个 tile 上执行，由程序员而非硬件调度器来分配工作。OpenAI 提到了 TensorInfo，这是一种显式编码布局（layout）的抽象。这很可能就是为 Jalapeño 设计的那套布局，将由 Linear Layouts 提供底层支持。最后，每个核心都提供数据预取（prefetching）和解耦的乱序执行单元。例如，用户可以对预取数据的等待进行编程，该等待通过信号量（semaphore）锁定。

In a weird twist of fate, OpenAI models like GPT 5.6 Sol, which currently run on NVIDIA GPUs, have been used to design a chip that poses a real threat to the CUDA moat - NVIDIA's own GPUs are helping usher in their potential successor in real time.

命运的诡异转折在于，OpenAI 的模型（如 GPT 5.6 Sol）目前运行在 NVIDIA GPU 上，却被用来设计一款对 CUDA 护城河构成切实威胁的芯片——NVIDIA 自家的 GPU 正在实时助推其潜在继任者的到来。

Comparing across time, we can also see Jalapeño's developmental pace, achieving more than 2x throughput improvements at certain interactivities in less than 2 weeks. Each tarball we get from the Jalapeño team has a world of wonders inside.

从时间维度对比，我们同样可以看到 Jalapeño 的开发节奏：不到两周内，在特定交互性（interactivity）水平下吞吐量提升超过 2 倍。我们从 Jalapeño 团队拿到的每一个 tarball，里面都藏着一个奇妙世界。

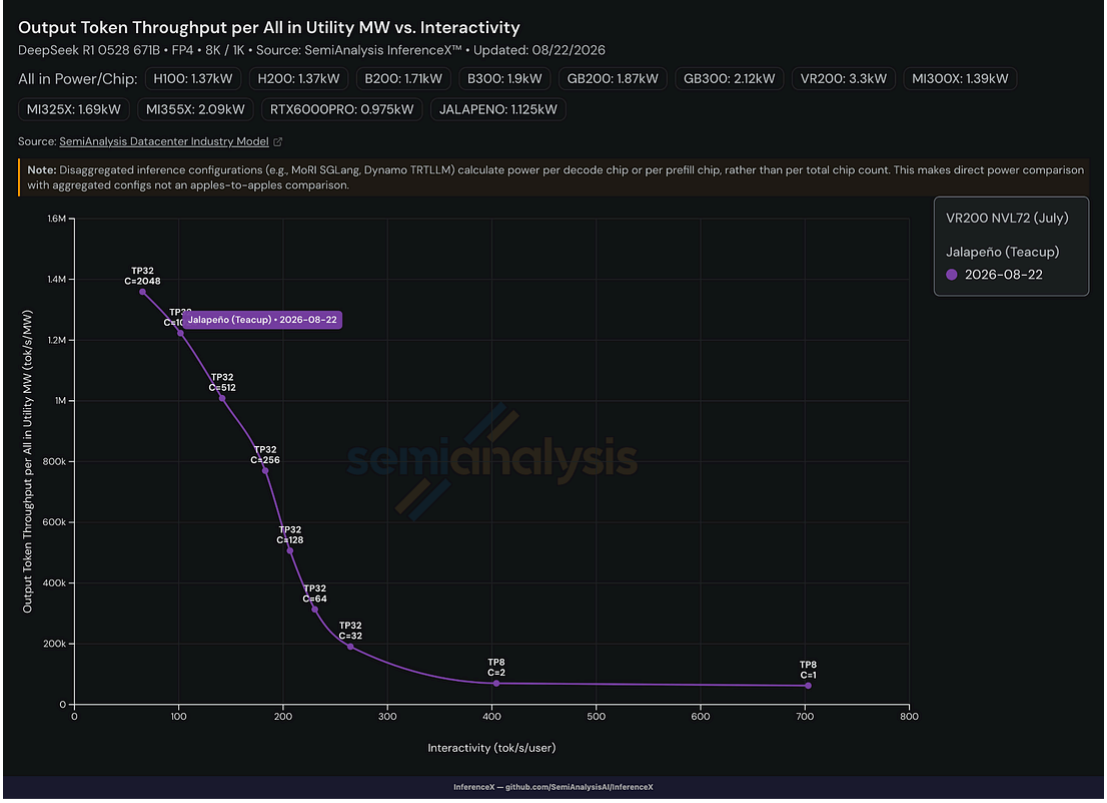


Source: OpenAI, SemiAnalysis

来源：OpenAI、SemiAnalysis

Not only did kernel performance improve, in the span of 8 days, the Jalapeño team enabled TP32, building on the previous TP8 configs and expanding beyond a single system to get a full rack-scale config running on a large model. This is a really impressive pace of development.

不仅内核性能有所提升，在 8 天时间内，Jalapeño 团队在先前 TP8 配置的基础上实现了 TP32，并突破单一系统限制，在大模型上跑通了完整的机架级（rack-scale）配置。这一开发速度确实令人印象深刻。



Source: OpenAI, SemiAnalysis

来源：OpenAI、SemiAnalysis

To validate performance before committing to real hardware runs, OpenAI also has a simulator “chilisim” accurate to within 5% of measured hardware, using a fixed-width trace bus. Tracing on A0 was limited but has improved substantially on B0, likely with inputs from actual runs on A0 silicon. Engineers have demoed the Codex CLI running an internal model, nicknamed “Raiku” or “5.3 Codex Spark”, at 1.2ms TPOT.

为了在投入真实硬件运行前验证性能，OpenAI 还开发了一款名为“chilisim”的模拟器，其精度与实测硬件误差在 5%以内，采用固定宽度 trace 总线。A0 芯片上的 trace 能力有限，但在 B0 上已有大幅改善，很可能得益于 A0 硅片实际运行数据的输入。工程师们已演示了 Codex CLI 运行一款内部模型（昵称“Raiku”或“5.3 Codex Spark”），TPOT 达到 1.2 毫秒。

The team also showed off Codex-written demos running directly on the chip: Doom at 36 FPS, an FP32 fluid-dynamics simulation, and a “Liquid Light” mouse-drag visualization.

团队还展示了由 Codex 编写、直接在芯片上运行的演示程序：Doom 以 36 FPS 运行、一个 FP32 流体动力学模拟，以及一个“Liquid Light”鼠标拖拽可视化效果。

Source: SemiAnalysis SemiAnalysis

On the model side, OpenAI's internal megakernel approach, nicknamed "gigakernel", is built around a single megakernel that loops on-device to reduce CPU overhead and launch time. The team is also leaning further into test-time compute strategies, with internal interest specifically in how to coherently use 1 million rollouts.

模型层面，OpenAI 内部采用的“megakernel”方案（内部昵称“gigakernel”）围绕单一巨型内核构建，通过在设备端循环执行来降低 CPU 开销和启动延迟。团队还在进一步加大测试时计算（test-time compute）策略的投入，内部尤其关注如何连贯地使用 100 万次 rollout。

To Disagg or Not to Disagg, 'tis the Question

去聚合（Disagg），还是不去聚合，这是个问题

We mentioned earlier that OpenAI is not using prefill decode disaggregation on these chips. This came as a surprise to us, as NVIDIA and AMD GPU performance benefits significantly from PDD, even on homogenous hardware. Let's dig into why the

Jalapeño team went this way.

我们此前提到，OpenAI 并未在这些芯片上采用 prefill/decode 分离（PDD）。这让我们颇感意外，因为即便在同类硬件上，NVIDIA 和 AMD GPU 的性能也能从 PDD 中显著受益。下面我们来深入探讨一下 Jalapeño 团队为何选择这条路线。

Prefill-decode disaggregation (PDD) looks attractive when the workload is frozen. Prefill and decode stress hardware differently, so assigning each phase to a separately tuned pool can improve efficiency at one chosen input/output ratio. Production traffic, however, does not stay at that ratio. Input and output sequence lengths, concurrency, cache-hit rates, speculative-acceptance rates, and latency targets all move throughout the day.

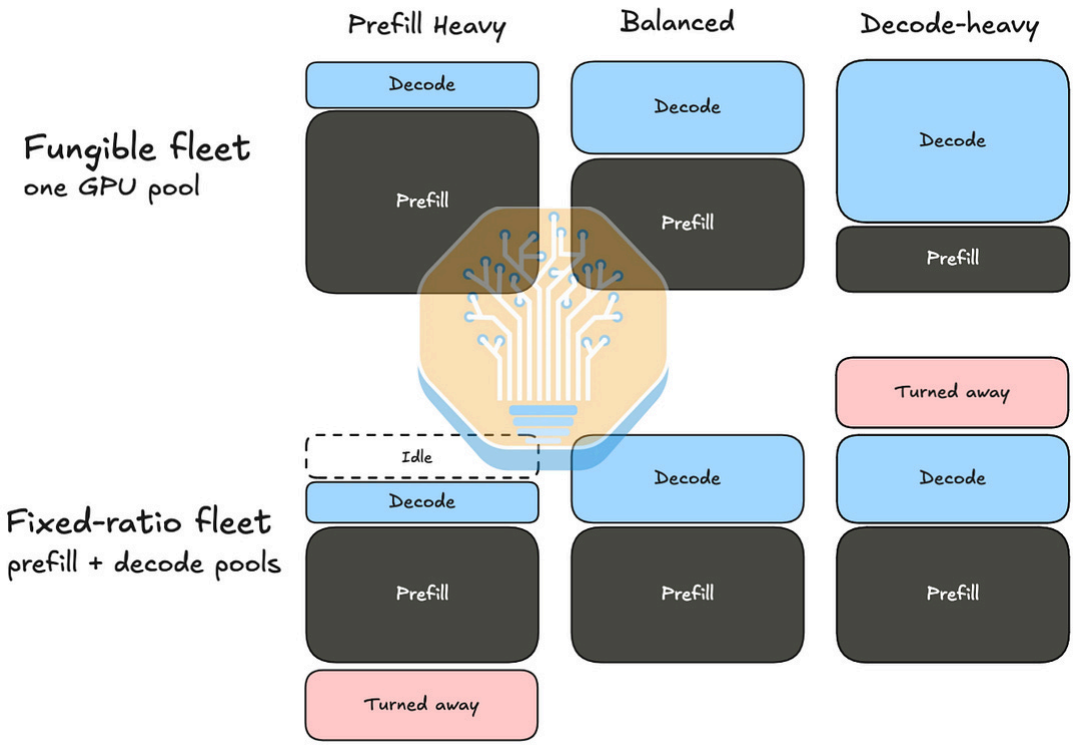
当工作负载固定不变时，预填充–解码解耦（PDD）颇具吸引力。预填充和解码对硬件的压力特征不同，因此将各阶段分配到独立调优的资源池，可以在某一选定的输入/输出比例下提升效率。然而，生产流量并不会维持在该比例。输入和输出序列长度、并发度、缓存命中率、投机接受率以及延迟目标，都会在一天中持续变化。

Once devices are divided into prefill and decode pools, too much prefill demand leaves decode chips idle while requests queue. But too much decode demand does the opposite. The operator must continuously predict the right split, provision spare capacity on both sides, and rebalance a system whose ideal ratio is always moving.

一旦设备被划分为预填充（prefill）和解码（decode）资源池，预填充需求过多时，解码芯片会闲置，而请求却在排队；反之，解码需求过多时则情况相反。运营方必须持续预测正确的拆分比例，在两侧预留冗余容量，并不断重新平衡一个理想配比始终在变动的系统。

In a unified system, some resources may be underused during a particular phase, but every device remains available to serve the next request. In a disaggregated system, an entire chip can sit idle simply because it belongs to the wrong pool. Local utilization looks better, but global utilization can be bad.

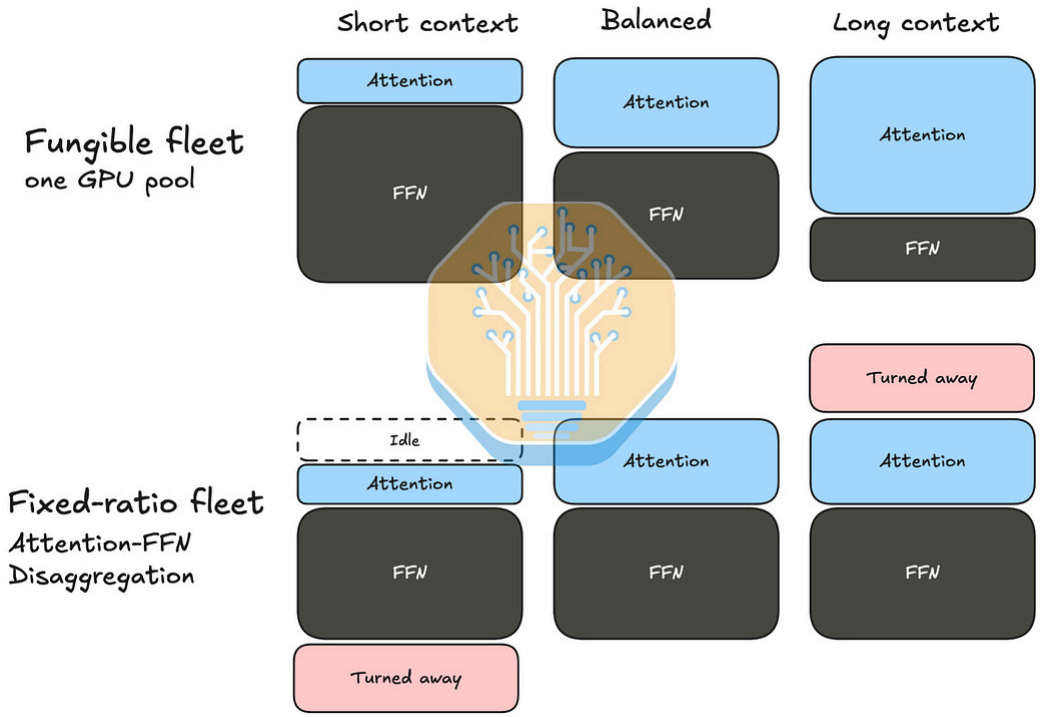
在统一系统中，某些资源在特定阶段可能利用不足，但每台设备仍可随时服务于下一个请求。而在解耦系统中，整颗芯片可能仅仅因为归属于错误的资源池而闲置。局部利用率看似良好，但全局利用率可能很差。



Source: SemiAnalysis SemiAnalysis

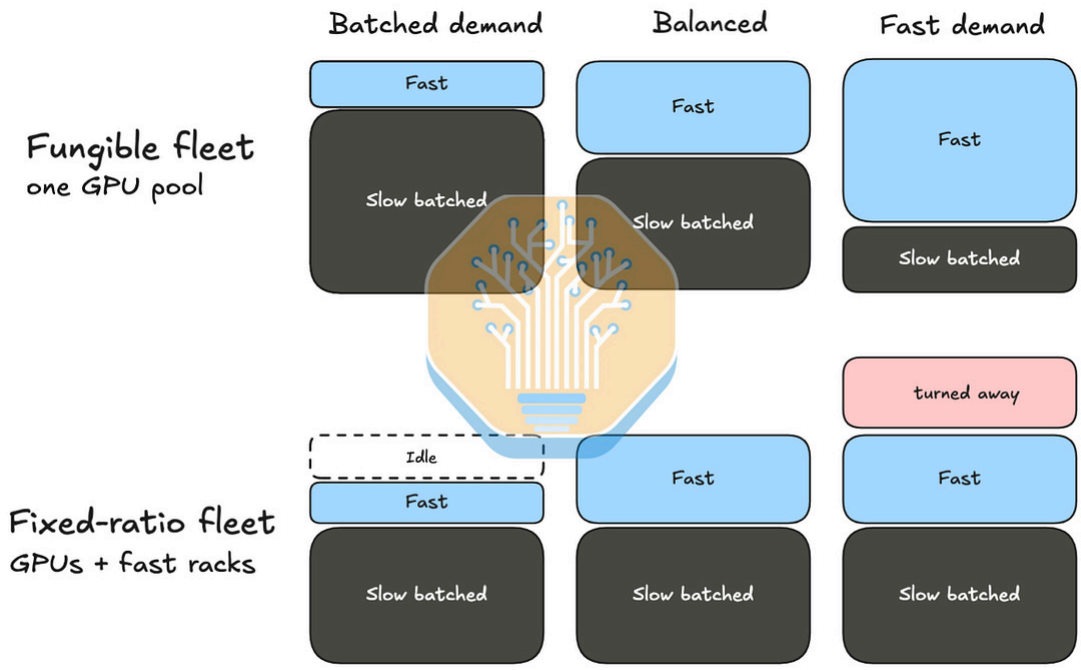
Disaggregation also breaks locality. The prefill worker produces a large KV cache that the decode worker immediately needs, so the system must transfer that state across the network before generation can continue. That adds bandwidth consumption, synchronization, queueing, and another failure domain. The cost also rises with input sequence length because KV cache grows. However, avoiding the movement of KVs is largely a power and latency optimization; being willing to move some KVs around can allow for increased hardware utilization at the expense of some power and per-request latency.

解耦架构还破坏了局部性。预填充（prefill）工作节点会产生大量 KV 缓存，而解码（decode）工作节点需要立即使用这些缓存，因此系统必须在生成继续之前通过网络传输该状态。这增加了带宽消耗、同步开销、排队延迟，并引入了新的故障域。同时，由于 KV 缓存随输入序列长度增长，相关成本也随之上升。然而，避免 KV 迁移在很大程度上是一种功耗与延迟优化；若愿意迁移部分 KV，则可在牺牲一定功耗和单请求延迟的前提下，换取更高的硬件利用率。



A fungible fleet shifts capacity between latency-sensitive requests and throughput-oriented batches, while a fixed split strands hardware whenever the traffic mix changes. Moreover, context length changes the balance between attention and FFN work, making any fixed hardware ratio efficient only near its design point.

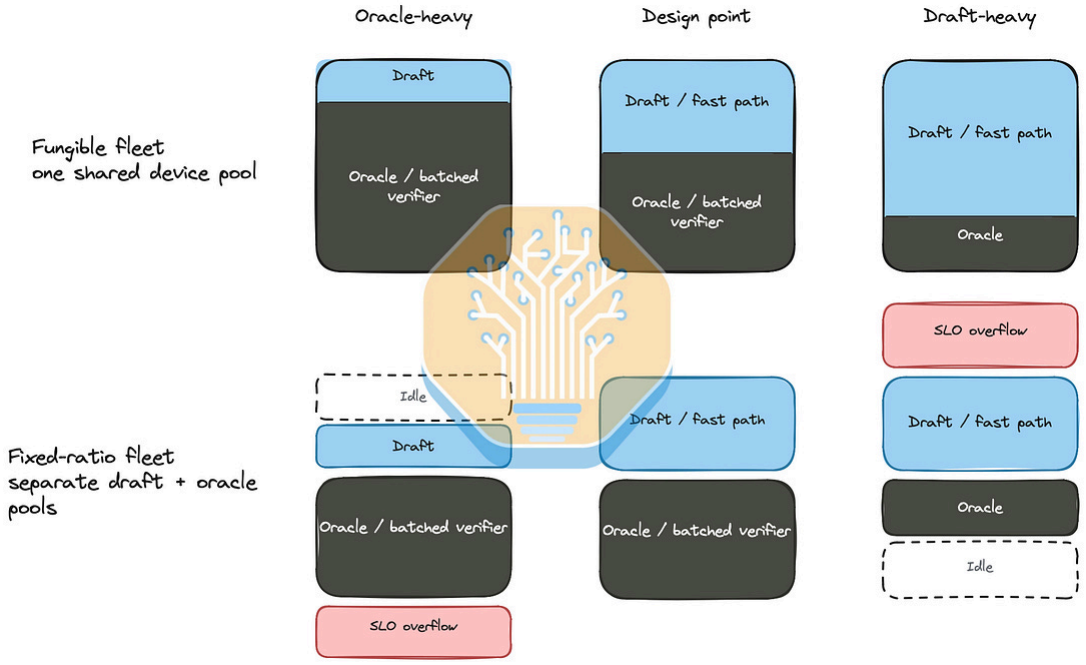
可灵活调度的算力池能在延迟敏感型请求与吞吐导向型批处理之间动态调配算力，而固定划分则会在流量结构变化时造成硬件闲置。此外，上下文长度会改变注意力计算与前馈网络（FFN）计算之间的配比，使得任何固定硬件配比仅在其设计点附近才保持高效。



Source: SemiAnalysis SemiAnalysis

The same constraint applies to speculative decoding. A draft model has to feed candidate tokens to the verifier with extremely low latency. Separating the two across specialized pools turns a tightly coupled decoding loop into a distributed protocol. The extra communication and coordination can consume the latency saved by drafting. Keeping both models on the same devices and low-latency fabric preserves the locality that makes speculation worthwhile in the first place.

同样的约束也适用于投机解码（speculative decoding）。草稿模型必须以极低延迟向验证器（verifier）提供候选 token。将两者分离到不同的专用资源池中，会把紧密耦合的解码循环变成分布式协议，额外的通信与协调开销可能抵消草稿生成所节省的延迟。将两个模型部署在同一设备及低延迟网络上，才能保持局部性（locality），而这正是投机解码之所以有价值的前提。



Source: SemiAnalysis SemiAnalysis

However, disaggregation can still win where demand is sufficiently large, stable, and predictable, particularly when conventional GPUs need large phase-specific batches to reach good throughput. But it is not free lunch.

然而，在需求足够庞大、稳定且可预测的情况下，分解（disaggregation）仍能胜出，尤其是当传统 GPU 需要大批量、分阶段的批次才能达到良好吞吐量时。但这并非免费的午餐。

From Japanese to Indian (Mild to Spicy): Katsu, Vindaloo, and Chana - How are the curry dishes put

together into a rack system?

从日式到印度式（微辣到重辣）：Katsu、Vindaloo 与 Chana——这些咖喱菜品如何整合进机架系统?

The Jalapeño System at the rack unit level consists of a CPU host rack and an ASIC rack. The host rack houses 16 host CPU trays named “Katsu,” each corresponding to one of the 16 ASIC trays, named “Vindaloo,” to the right of the Katsu. Each host houses two Turin-class AMD EPYC CPUs with 1.5TB of DRAM, 2x E1.S, and 2x M.2 SSDs per rack. Each tray is also specced with 400G (2x200G) frontend networking. Each Katsu tray connects to each Vindaloo tray via 8 external PCIe DAC cables that run horizontally across the rack at the front. The system level design is done in partnership with Celestica.

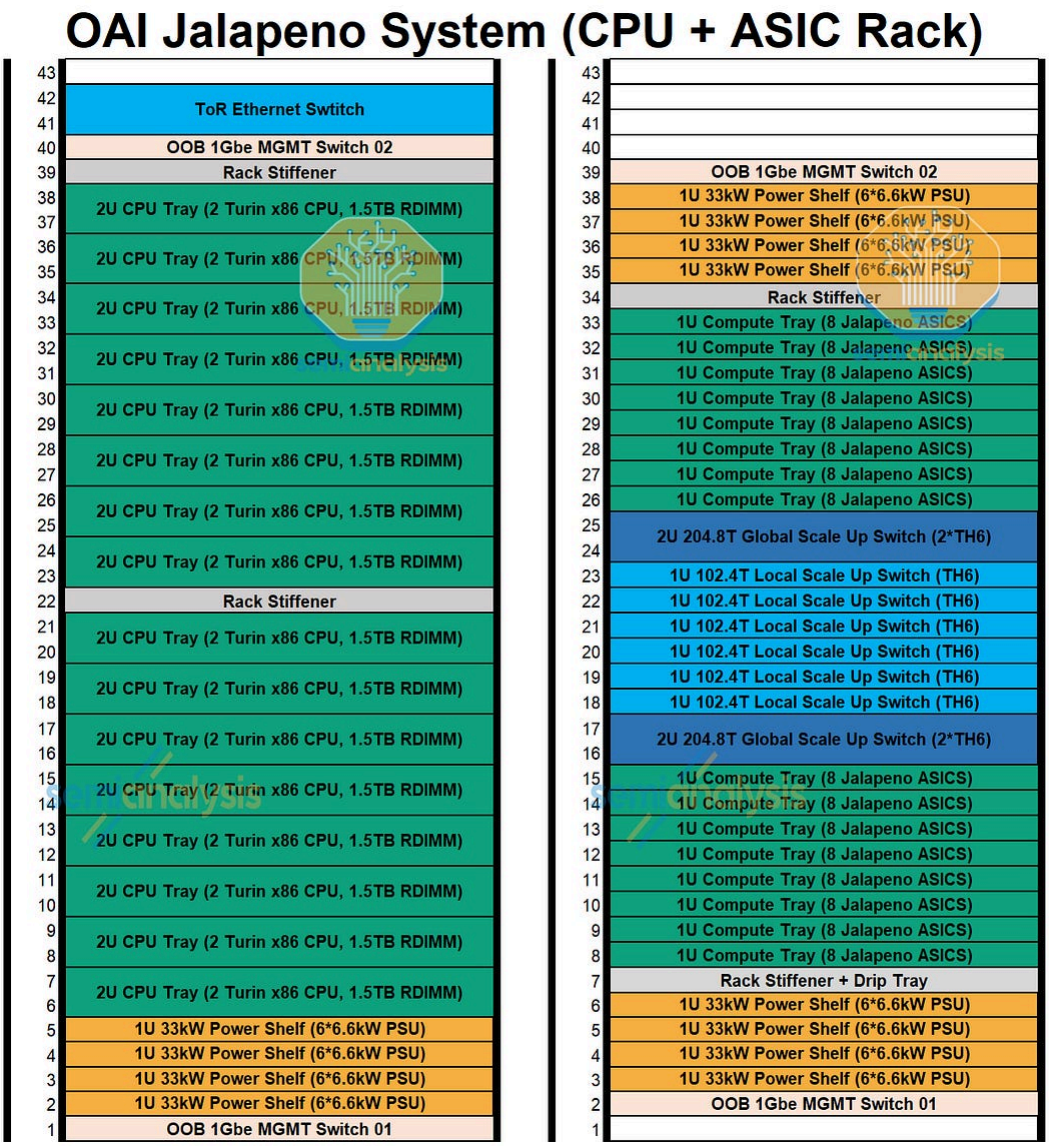
Jalapeño 系统在机架单元层面由 CPU 主机机架和 ASIC 机架组成。主机机架容纳 16 个名为“Katsu”的主机 CPU 托盘，每个托盘对应右侧 16 个名为“Vindaloo”的 ASIC 托盘之一。每台主机搭载两颗 Turin 级 AMD EPYC CPU，配备 1.5TB DRAM、2x E1.S 和 2x M.2 SSD（每机架）。每个托盘还标配 400G（2x200G）前端网络。每个 Katsu 托盘通过 8 根外部 PCIe DAC 线缆与每个 Vindaloo 托盘连接，这些线缆在机架前部水平横跨。系统级设计与 Celestica 合作完成。

The ASIC rack consists of 16 Vindaloo trays and 8 scale up switch trays (6 for local + 2 for global), named “Chana.” Each Vindaloo tray consists of 8 Jalapeño ASICs, making up a total of 128 Jalapeño ASICs per rack. The ASICs are connected to each of the Chana switch trays via a copper cable backplane, just like that of Nvidia’s Oberon. The scale up topology is split into a local domain of 128 ASICs within the rack and a global domain connecting up to 16 racks or 2,048 ASICs. We will explain the bandwidth and the topology in more detail below.

该 ASIC 机架由 16 个 Vindaloo 托盘和 8 个 scale-up 交换机托盘（6 个用于本地，2 个用于全局）组成，命名为“Chana”。每个 Vindaloo 托盘包含 8 个 Jalapeño ASIC，因此每个机架共有 128 个 Jalapeño ASIC。ASIC 通过铜缆背板连接到每个 Chana 交换机托盘，这与 NVIDIA 的 Oberon 类似。Scale-up 拓扑分为机架内 128 个 ASIC 的本地域，以及连接最多 16 个机架或 2,048 个 ASIC 的全局域。我们将在下文更详细地解释带宽和拓扑结构。

Power provisioning to a sidecar host rack draws roughly 50kW provisioned (31kW in production), and the ASIC rack draws 130kW, making the total two rack system roughly 160kW. That's basically a double-wide GB300 rack in terms of power draw.

侧车主机机架的供电配置约为 50kW（实际生产功耗 31kW），而 ASIC 机架功耗为 130kW，因此整个双机架系统总功耗约为 160kW。就功耗而言，这基本上相当于一个双宽 GB300 机架。



Source: SemiAnalysis, OpenAI

来源：SemiAnalysis，OpenAI

OpenAI can connect up to 2,048 Jalapeño XPUs within a single scale-up network. The scale-up network consists of two domains, a local domain connecting all 128 XPUs over backplane within the rack, as well as a global domain connecting 2,048 XPUs over 16 racks using a hybrid of copper and optical interconnect. Each rack consists of 8 Chana switch trays. Six Chana switches in the middle are for the local domain, which come with one 102.4T Tomahawk 6 switch ASIC each. Two Chana switches at the top and bottom of the local switches are for the global domain, which we think could consist of 2x 102.4T Tomahawk 6 switches making up to 204.8T per

switch tray.

OpenAI 可在单个 scale-up 网络内连接最多 2,048 个 Jalapeño XPU。该 scale-up 网络由两个域组成：本地域通过机架内背板连接全部 128 个 XPU；全局域则通过 16 个机架、采用铜缆与光互联混合方案连接 2,048 个 XPU。每个机架包含 8 个 Chana 交换托盘。中间 6 个 Chana 交换机用于本地域，每个配备一颗 102.4T Tomahawk 6 交换 ASIC。本地交换机顶部和底部的两个 Chana 交换机用于全局域，我们认为其可能由 2 颗 102.4T Tomahawk 6 交换机组成，每个交换托盘最高可达 204.8T。

In the local domain, each of the 128 Jalapeño chips has a per XPU uni-directional bandwidth of 4.8Tb/s and is connected on an all-to-all basis to 6x 102.4Tb/s Tomahawk 6 ASICs. This would amount to 48-differential pair (DP) male and female connector pairs per XPU translating to a total of 6,144DPs worth of passive copper cables per rack used for local scale-up.

在本地域（local domain）中，128 颗 Jalapeño 芯片中的每一颗均具备每 XPU 4.8Tb/s 的单向带宽，并以全互联（all-to-all）方式连接至 6 颗 102.4Tb/s Tomahawk 6 ASIC。这相当于每个 XPU 对应 48 对差分信号（DP）公母连接器，合计每机架使用 6,144 对 DP 的无源铜缆用于本地 scale-up 互联。

For the global domain, 16 racks totaling 2,048 XPUs are connected together via a combination of copper backplane, electrical 204.8T TH6 switch, 1.6T transceivers, and optical circuit switch. Each XPU has a uni-directional bandwidth of 1.6Tb/s for the global link, which is 16-differential pair (DP) male and female connector pairs per XPU for the backplane between the XPU and the global switch. Bandwidth exiting each global switch tray of 2 ASICs each is split between the backplane and front panel optics.

在全局域（global domain）中，16 个机架共计 2,048 颗 XPU 通过铜背板、电信号 204.8T TH6 交换机、1.6T 光收发器及光路交换机（optical circuit switch）组合互联。每颗 XPU 的全局链路单向带宽为 1.6Tb/s，即 XPU 与全局交换机之间背板对应每 XPU 16 对差分信号（DP）公母连接器。每个全局交换机托盘（含 2 颗 ASIC）的出口带宽在背板与前板光模块之间分配。

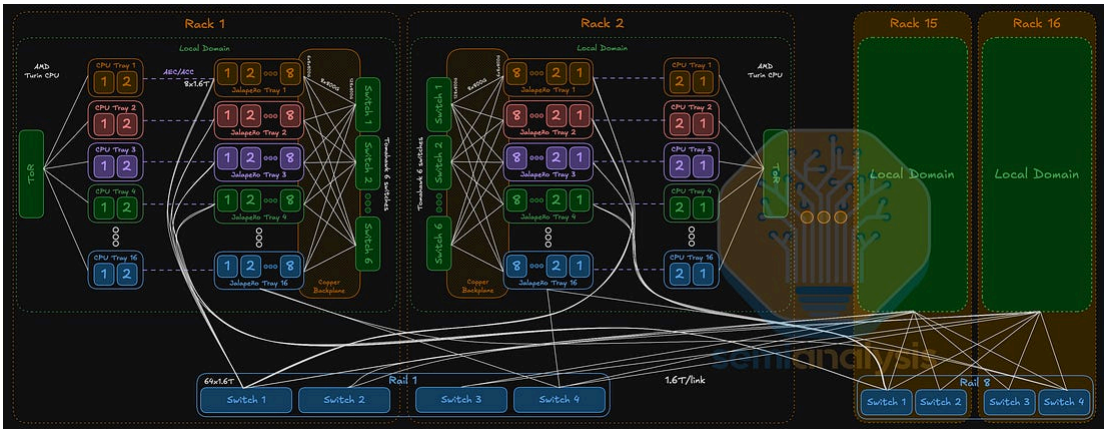
Between local domain and global domain, backplane connector count per rack comes up to 64 DPs per XPU and a total of 8,192 DPs worth of passive copper cables per

rack.

本地域与全域之间，每个机架的背板连接器数量达到每 XPU 64 个 DP，全机架合计 8,192 个 DP 的无源铜缆。

The global domain adopts a rail-only architecture consisting of 8-rails across the global domain. We think OpenAI routes optical links in the global domain via Optical Circuit Switches (OCS) installed in every rack. For every XPU, 1.6Tb/s of global bandwidth will travel to the global switch tray over the copper backplane. This then exits the switch through the front panel via 1.6T transceivers, which go to the passive optical switch before exiting the rack. This expands the scale-up world size to 2,048 XPUs combining 16 racks of 128 XPUs each.

全域采用纯 rail 架构，整个全域共 8 条 rail。我们认为 OpenAI 通过每个机架中安装的光路交换机（OCS）来路由全域光链路。每个 XPU 的 1.6Tb/s 全域带宽将通过铜背板传输至全域交换托盘，随后经前面板通过 1.6T 光收发器输出，进入无源光交换机后再出机架。这将 scale-up 世界规模扩展至 2,048 个 XPU，由 16 个机架组成，每个机架 128 个 XPU。



Source: SemiAnalysis SemiAnalysis

Because scale-up networking is only about 10% of total system cost, that flexibility buys valuable optionality for future 10–20 trillion parameter models or 2–4 million token context windows. On deployment, OpenAI is partnering with neoclouds and is gathering reliability data with datacenter partners through January while optimizing dock-to-rack rollout time.

由于 scale-up 网络仅占系统总成本约 10%，这种灵活性为未来 10–20 万亿参数模型或 2–4 百万 token 上下文窗口提供了宝贵的期权价值。在部署方面，OpenAI 正与 neocloud 合作，并在 1 月前与数据中心合作伙伴收集可靠性数据，同时优化从到货到上架（dock-to-rack）的部署时间。

What's Next 接下来是什么

Next, we talk about the future of Jalapeño, whose first production token is coming soon. The next goal is 100MW, and the hurdles will mostly be hardware: How much can they produce, how well can they deploy and operate datacenters, how do they handle monitoring, and resiliency, etc. The software is already proven, and with internal models, every software headstart is easily caught up to. Behind the paywall we will discuss implications for NVIDIA, AMD, Cerebras, and other chip companies who have signed deals with OpenAI in the coming years.

接下来我们谈谈 Jalapeño 的未来，其首个生产级 token 即将推出。下一个目标是 100MW，主要障碍将集中在硬件层面：产能有多大，数据中心部署和运营能力如何，监控与容灾怎么处理，等等。软件已经得到验证，而且凭借内部模型，软件方面的先发优势很容易被追平。付费墙之后，我们将讨论这对 NVIDIA、AMD、Cerebras 以及其他已与 OpenAI 签署未来数年合作协议的芯片公司意味着什么。

We also cover production volumes, units, and timelines for the next generation chip in our Accelerator model here.

我们在本地的 Accelerator 模型中也涵盖了下一代芯片的产量、出货量及时间线。

Direct Implications for NVIDIA, AMD and Cerebras...

对 NVIDIA、AMD 和 Cerebras 的直接影响……