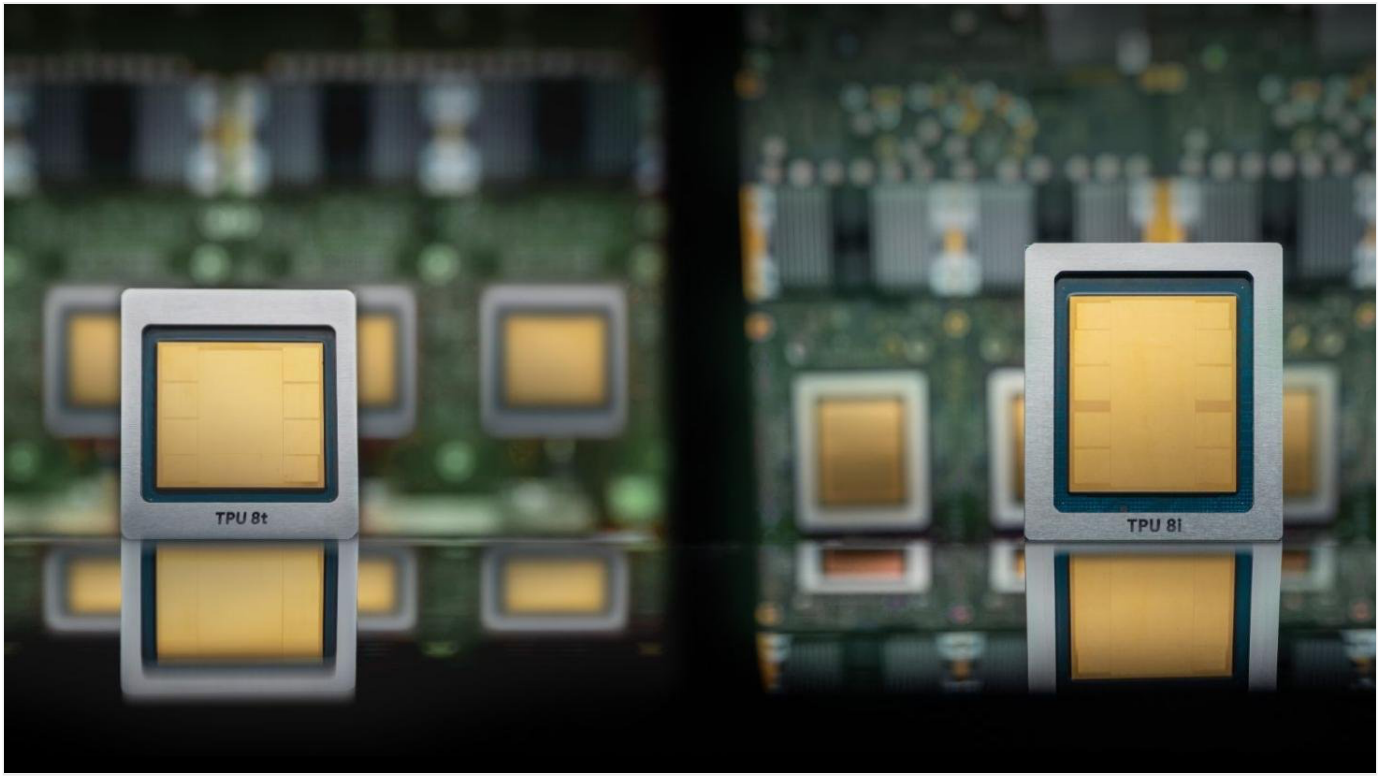
Compute

Inside the eighth-generation TPU: An architecture deep dive

深入了解第八代 TPU：架构深度解析

April 23, 2026



Diwakar Gupta
Distinguished Engineer, Google Cloud

Google Cloud 杰出工程师

Sabastian Mugazambi
Group Product Manager, Google Cloud

Google Cloud 集团产品经理

At Google, our TPU design philosophy has always been centered on three pillars: scalability, reliability, and efficiency.

在 Google，我们的 TPU 设计理念始终围绕着三大支柱：可扩展性、可靠性和效率。

As AI models evolve from dense large language models (LLMs) to massive Mixture-of-Experts (MoEs) and reasoning-heavy architectures, the hardware must do more than just add floating point operations per second (FLOPS); it must evolve to meet the specific operational intensities of the latest workloads.

随着 AI 模型从稠密的大型语言模型 (LLMs) 演进为海量的混合专家模型 (MoEs) 以及重推理架构，硬件不仅需要增加每秒浮点运算次数 (FLOPS)，还必须不断进化，以满足最新工作负载特定的运算强度。

The rise of agentic AI requires infrastructure that can handle long context windows and complex sequential logic.

代理式 AI (Agentic AI) 的兴起要求基础设施能够处理长上下文窗口和复杂的顺序逻辑。

At the same time, world models have emerged as a necessary evolution from current next-sequence-of-data architectures, which means newer agents are simulating future scenarios, anticipating consequences, and learning through "imagination" rather than risky trial-and-error.

与此同时，世界模型已成为当前“下一序列数据”架构的必然演进方向，这意味着新一代代理正在模拟未来场景、预判后果，并通过“想象”而非冒险的试错来进行学习。

The eighth-generation TPUs (TPU 8t and TPU 8i) are our answer to these challenges, ensuring that every workload, from the first token of training to the final step of a multi-turn reasoning chain, is running on the most efficient path possible.

第八代 TPU（TPU 8t 和 TPU 8i）是我们应对这些挑战的方案，

确保从训练的第一个 token 到多轮推理链的最后一步，每一项工作负载都能在最高效的路径上运行。

They are built to efficiently train and serve world models like Google DeepMind's Genie 3, enabling millions of agents to practice and refine their reasoning in diverse simulated environments.

它们旨在高效地训练和提供像 Google DeepMind 的 Genie 3 这样的世界模型，使数百万个智能体能够在各种模拟环境中练习并完善其推理能力。

TPU 8: Specialized by design

TPU 8：专为特定设计而生

Recognizing that the infrastructure requirements for pre-training, post-training, and real-time serving have diverged, our eighth-generation TPUs introduce two distinct systems: TPU 8t and TPU 8i.

意识到预训练、后训练和实时推理对基础设施的需求已各不相同，我们的第八代 TPU 推出了两个截然不同的系统：TPU 8t 和 TPU 8i。

These new systems are key components of Google Cloud's AI Hypercomputer, an integrated supercomputing architecture that combines hardware, software, and networking to power the full AI lifecycle.

这些新系统是 Google Cloud AI 超级计算机（AI Hypercomputer）的关键组成部分，这是一种集成了硬件、软件

Hypercomputer 的关键组成部分。这是一个集成了硬件、软件和网络的超级计算架构，旨在为整个 AI 生命周期提供动力。

While both systems share the core DNA of Google's AI stack and support the full AI lifecycle, each is built to address distinct bottlenecks and optimize efficiency for critical stages of development.

虽然这两个系统都承袭了 Google AI 技术栈的核心基因，并支持完整的 AI 生命周期，但它们各自的构建目标是解决不同的瓶颈，并针对开发的关键阶段优化效率。

Additionally, by integrating Arm-based Axion CPU headers across our eighth-generation TPU system, we've removed the host bottleneck caused by data preparation latency.

此外，通过在第八代 TPU 系统中集成基于 Arm 架构的 Axion CPU 节点，我们消除了由数据准备延迟引起的主机瓶颈。

Axion provides the compute headroom to handle complex data preprocessing and orchestration, so that TPUs stay fed and don't stall.

Axion 提供了充足的计算余量来处理复杂的数据预处理和编排，从而确保 TPU 能够持续获得数据输入而不发生停顿。

TPU 8t: The pre-training powerhouse

TPU 8t：预训练的动力源泉

Optimized for massive-scale pre-training and embedding-heavy workloads, TPU 8t utilizes our proven 3D torus network topology at an even larger scale of 9,600 chips in a single superpod. TPU 8t is designed for maximum throughput across hundreds of superpods

designed for maximum throughput across hundreds of superpods, ensuring that training runs stay on schedule.

TPU 8t 专为大规模预训练和高嵌入负载（embedding-heavy workloads）而优化，采用了我们成熟的 3D 环面（torus）网络拓扑结构，在单个超算集群（superpod）中实现了高达 9,600 个芯片的更大规模。TPU 8t 旨在跨数百个超算集群实现最大吞吐量，确保训练任务按计划进行。

Here are some key advancements of TPU 8t over prior-generation TPUs:

以下是 TPU 8t 相比前代 TPU 的一些关键进步：

- **The SparseCore advantage**: Central to TPU 8t is the SparseCore, a specialized accelerator designed to handle the irregular memory access patterns of embedding lookups.

  SparseCore 优势：TPU 8t 的核心是 SparseCore，这是一种专门设计的加速器，用于处理嵌入查找（embedding lookups）中不规则的内存访问模式。

  While the Matrix Multiply Unit (MXU) handles matrix math, the SparseCore offloads data-dependent all-gather operations, amongst other collectives, preventing the zero-op bottlenecks that often plague general-purpose chips.

  虽然矩阵乘法单元 (MXU) 负责处理矩阵运算，但 SparseCore 承担了数据依赖的 all-gather 操作以及其他集合通信任务，从而防止了经常困扰通用芯片的零操作 (zero-op) 瓶颈。

- **VPU/MXU overlap and balanced scaling**: TPU 8t is designed to

- **VPU/MXU overlap and balanced scaling**: TPU 8t is designed to maximize provisioned FLOPs utilization. By implementing more balanced Vector Processing Unit (VPU) scaling, the architecture minimizes exposed vector operation time.

  VPU/MXU 重叠与平衡缩放：TPU 8t 旨在最大化已配置 FLOPs 的利用率。通过实现更平衡的向量处理单元 (VPU) 缩放，该架构最大限度地减少了暴露的向量操作时间。

  This allows for better overlapping of quantization, softmax, and layernorms with the matrix multiplications in the MXU, helping the chip stay busy rather than waiting on sequential vector tasks.

  这使得量化、softmax 和层归一化 (layernorms) 能够更好地与 MXU 中的矩阵乘法重叠，从而帮助芯片保持繁忙状态，而不是等待顺序向量任务。

- **Native FP4**: TPU 8t introduces native 4-bit floating point (FP4) to overcome memory bandwidth bottlenecks, doubling MXU throughput while maintaining accuracy for large models even at lower-precision quantization.

  原生 FP4：TPU 8t 引入了原生 4 位浮点 (FP4) 以克服内存带宽瓶颈，在将 MXU 吞吐量翻倍的同时，即使在低精度量化下也能保持大模型的准确性。

  By reducing the bits per parameter, the platform minimizes energy-intensive data movement and allows larger model layers to fit within local hardware buffers for peak compute utilization.

  通过减少每个参数的位数，该平台最大限度地降低了高能耗的数据传输，并允许更大的模型层驻留在本地硬件缓冲区内，从

数据传输，并允许更大的模型层容纳在本地软件缓存区内，从而实现峰值计算利用率。

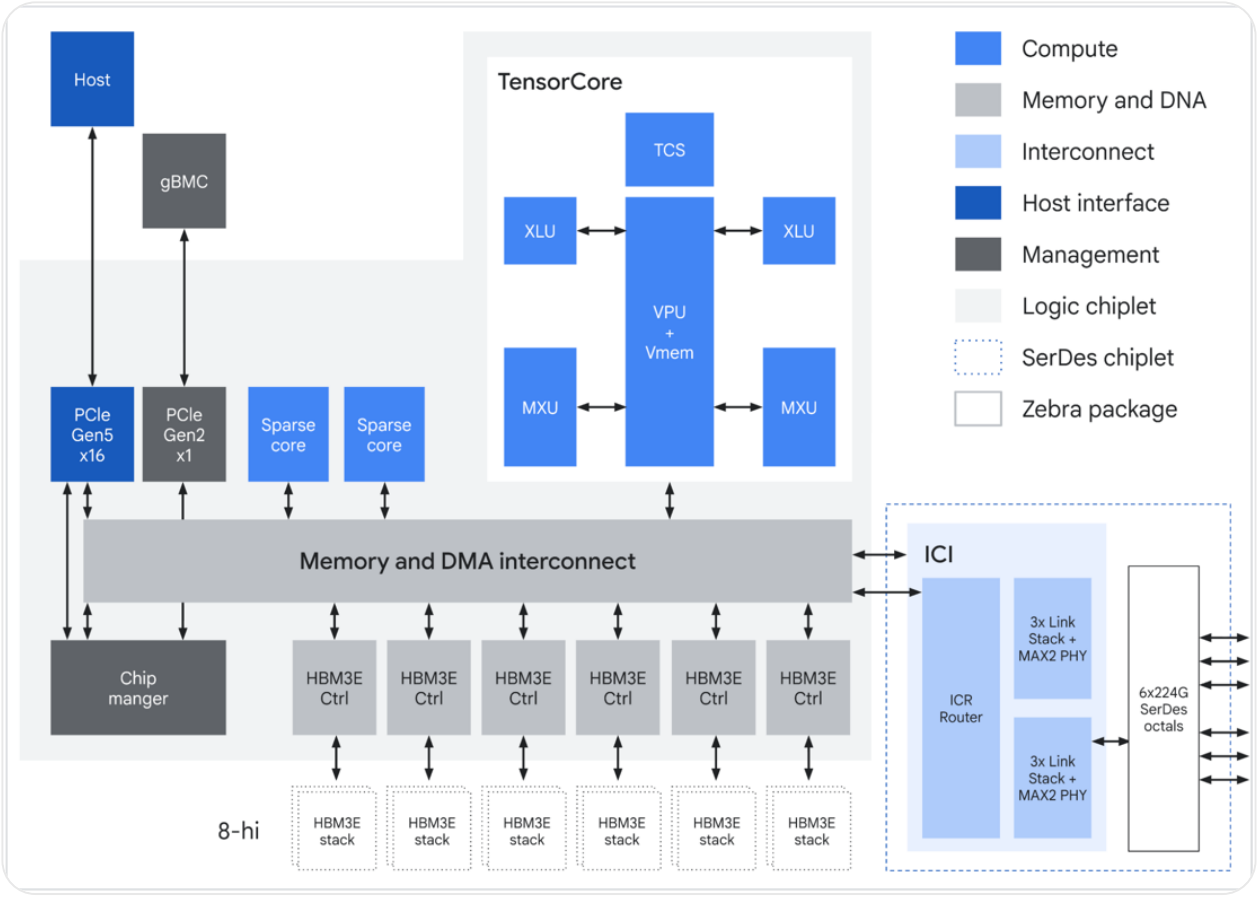


Figure 1: TPU 8t ASIC block diagram

图 1：TPU 8t ASIC 框图

- **Virgo Network topology and up to 4x data center network increase**: To support the massive data requirements of TPU 8t, we introduced Virgo Network. This new networking architecture enables up to 4x increased data center network (DCN) bandwidth on TPU 8t training over DCN. Virgo Network is a scale-out fabric designed for the extreme requirements of modern AI workloads.

Virgo 网络拓扑及高达 4 倍的数据中心网络提升：为了支持 TPU 8t 海量的数据需求，我们引入了 Virgo 网络。这种全新的网络架构使 TPU 8t 在数据中心网络（DCN）上的训练带宽提升了高达 4 倍。Virgo 网络是一种专为现代 AI 工作负载的极端需求而设计的横向扩展架构。

Built on high-radix switches that reduce network layers by allowing more ports per switch, it employs a flat, two-layer non-blocking topology. Compared with traditional datacenter networks, this significantly reduces latency by minimizing network tiers. It features a multi-planar design with independent control domains to connect TPU 8t chips.

它基于高基数（high-radix）交换机构建，通过允许每个交换机拥有更多端口来减少网络层级，并采用扁平的双层无阻塞拓扑结构。与传统的数据中心网络相比，这通过最大限度地减少网络层级显著降低了延迟。它采用多平面设计，具有独立的控制域来连接 TPU 8t 芯片。

The TPU 8t racks also connect with the Jupiter north-south fabric to access compute and storage services. Together, this streamlined architecture delivers the massive bisection bandwidth and deterministic low latency necessary for enabling the world's largest training clusters with high availability.

TPU 8t 机架还与 Jupiter 南北向网络架构相连，以访问计算和存储服务。这种精简的架构共同提供了海量的对分带宽和确定性的低延迟，这对于构建具有高可用性的全球最大规模训练集群至关重要。

With 2x scale-up bandwidth on the inter-chip interconnect (ICI) and up to 4x raw scale-out DCN bandwidth compared to the previous generation, TPU 8t drastically reduces data bottlenecks. Then, to further accelerate the development of frontier models, we scale distributed training beyond a single cluster. With JAX and Pathways, **we can now scale to more than 1 million TPU chips in a single training cluster**. Virgo Network can link over 134,000 TPU 8t chips with up to 47 petabits/sec of non-blocking bi-sectional bandwidth in a single fabric. This fabric delivers over 1.6 million ExaFlops with near-linear scaling performance.

与上一代相比，TPU 8t 的芯片间互连（ICI）扩展带宽提升了 2 倍，原始扩展（scale-out）DCN 带宽提升了高达 4 倍，从而大幅减少了数据瓶颈。此外，为了进一步加速前沿模型的开发，我们将分布式训练扩展到了单个集群之外。借助 JAX 和 Pathways，我们现在可以在单个训练集群中扩展到超过 100 万个 TPU 芯片。Virgo 网络可以在单个架构中连接超过 134,000 个 TPU 8t 芯片，提供高达 47 PB/s 的无阻塞对分带宽。该架构可提供超过 160 万 ExaFlops 的算力，并具有近乎线性的扩展性能。

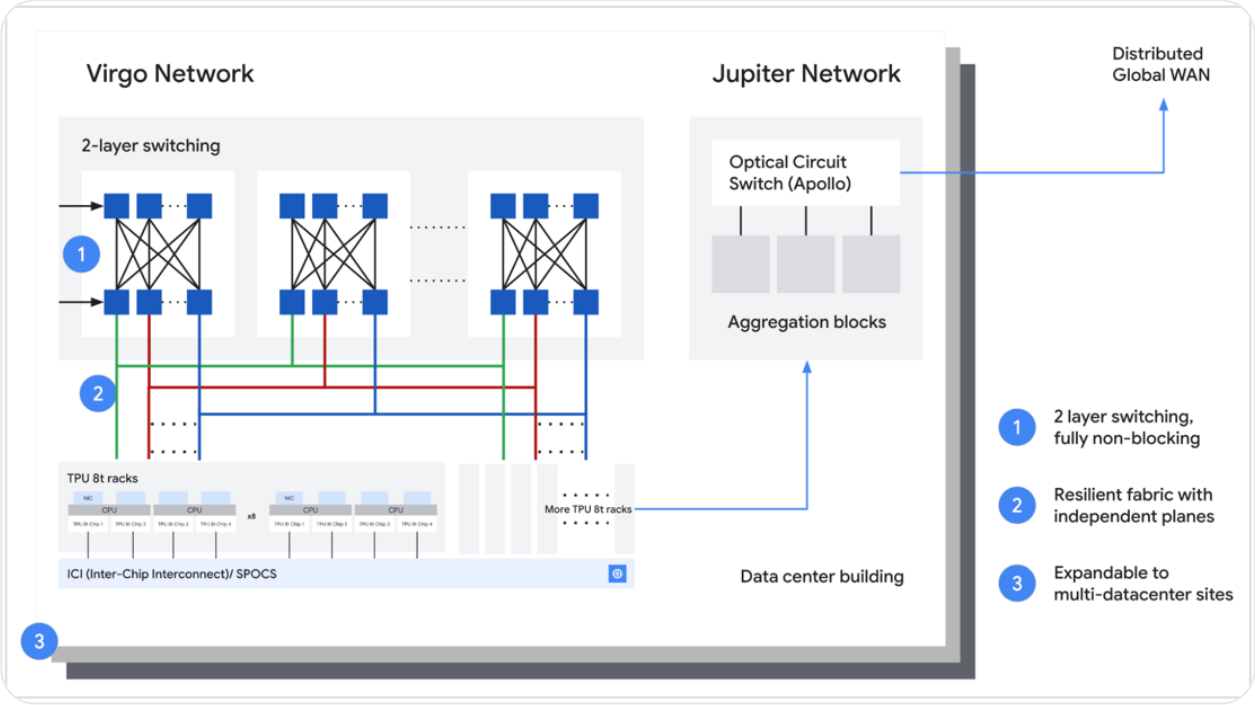


Figure 2: TPU 8t rack level connectivity to Virgo fabric

图 2：TPU 8t 机架级与 Virgo 架构的连接

- ***Faster storage access:*** *We are introducing* ***TPUDirect RDMA*** *and* ***TPU Direct Storage*** *in TPU 8t. TPU Direct RDMA enables direct*

data transfers between the TPU's memory (HBM) and the Network Interface Cards (NICs), bypassing the host CPU and DRAM. This reduces latency and host system bottlenecks, increasing the effective bandwidth for TPU-to-TPU communication.

> ***更快的存储访问：我们在 TPU 8t 中引入了 TPUDirect RDMA 和 TPU Direct Storage。TPU Direct RDMA 实现了 TPU 内存（HBM）与网络接口卡（NIC）之间的直接数据传输，绕过了主机 CPU 和 DRAM。这降低了延迟和主机系统瓶颈，提高了 TPU 间通信的有效带宽。***

Similarly, TPUDirect Storage bypasses CPU host bottlenecks by

enabling direct memory access between the TPU and high-speed managed storage like 10T Lustre, effectively doubling the bandwidth for massive data transfers.

> ***同样，TPUDirect Storage 通过在 TPU 与 10T Lustre 等高速托管存储之间实现直接内存访问，绕过了 CPU 主机瓶颈，从而使海量数据传输的带宽有效翻倍。***

This architecture allows the silicon to ingest training data at line rate, ensuring that the MXUs stay fully saturated even when processing large multimodal datasets.

> ***这种架构允许芯片以线速摄取训练数据，确保即使在处理大型多模态数据集时，MXU 也能保持完全饱和。***

By combining Managed Lustre 10T and TPUDirect Storage to route hundred-petabyte datasets directly to the silicon, TPU 8t prevents training delays caused by data ingestion bottlenecks. This delivers 10x faster storage access compared to training on seventh-generation

Ironwood TPUs.

通过结合 Managed Lustre 10T 和 TPUDirect Storage 将百 PB 级的数据集直接路由到芯片，TPU 8t 防止了由数据摄取瓶颈引起的训练延迟。与在第七代 Ironwood TPU 上进行训练相比，这提供了快 10 倍的存储访问速度。

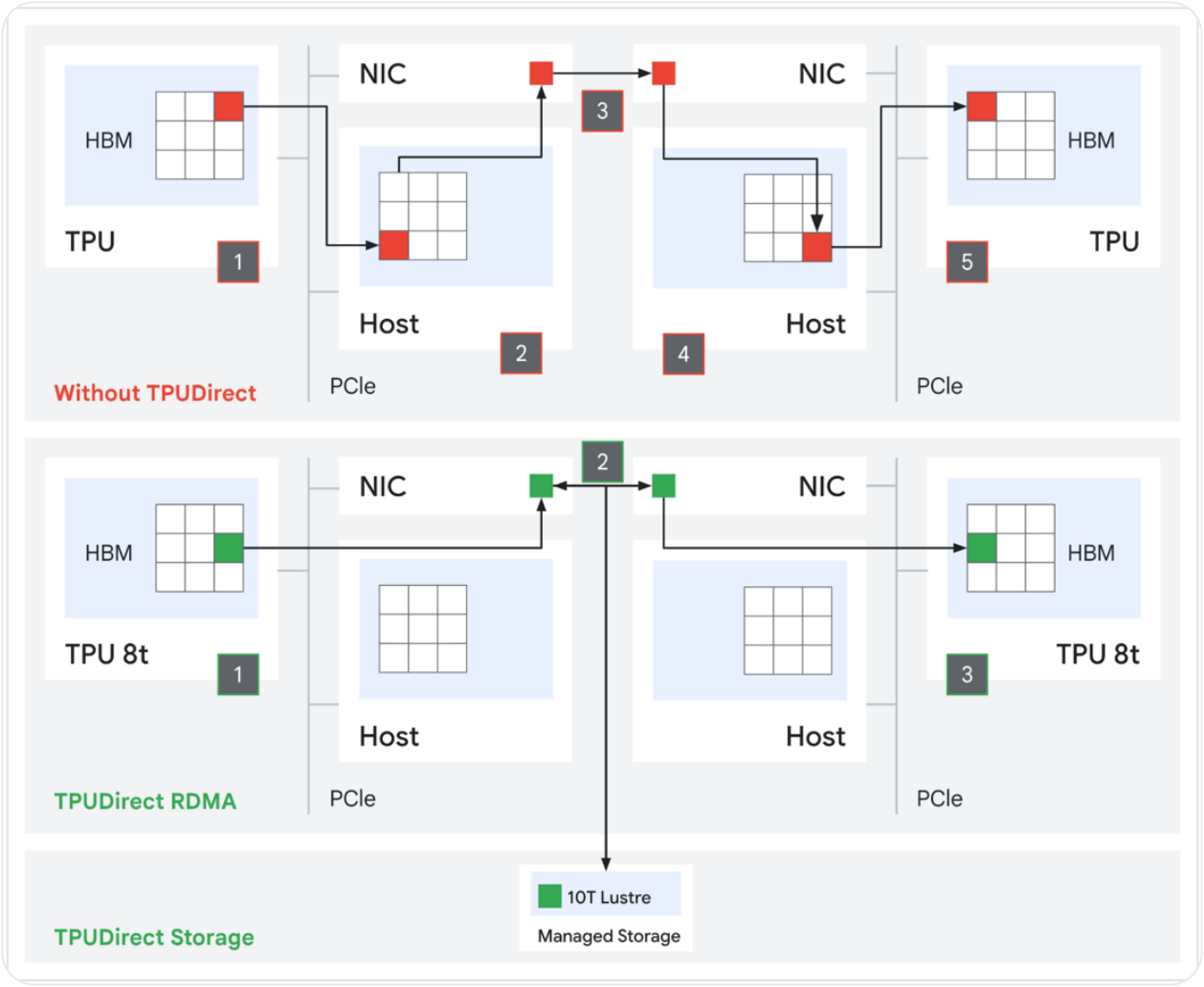


Figure 3: The top diagram shows the data transfer path without TPUDirect Storage. The bottom diagram shows TPU 8t data transfer with TPUDirect

Storage between 2 TPU 8t chips and TPUDirect Storage with Managed 10T Lustre storage.

图 3：上方图表显示了没有 TPUDirect Storage 的数据传输路径。下方图表显示了 2 个 TPU 8t 芯片之间使用 TPUDirect Storage 的数据传输，以及配合 Managed 10T Lustre 存储使用 TPUDirect Storage 的情况。

TPU 8i: The sampling and serving specialist

TPU 8i：采样与推理服务专家

Optimized for post-training and high-concurrency reasoning, we designed TPU 8i with our highest on-chip SRAM, a new Collectives Acceleration Engine (CAE), and a new serving-optimized network topology called Boardfly.

TPU 8i 专为训练后处理和高并发推理而优化。我们在设计中为其配备了容量最高的片上 SRAM、全新的集合通信加速引擎（CAE），以及一种名为 Boardfly 的新型推理服务优化网络拓扑。

- **Large on-chip SRAM:** With 3x more on-chip SRAM over the previous generation, TPU 8i can host a larger KV Cache entirely on silicon, significantly reducing the idle time of the cores during long-context decoding.

  大容量片上 SRAM：TPU 8i 的片上 SRAM 容量是上一代的 3 倍，可以直接在芯片上托管更大的 KV Cache，从而显著减少长上下文解码过程中核心的空闲时间。

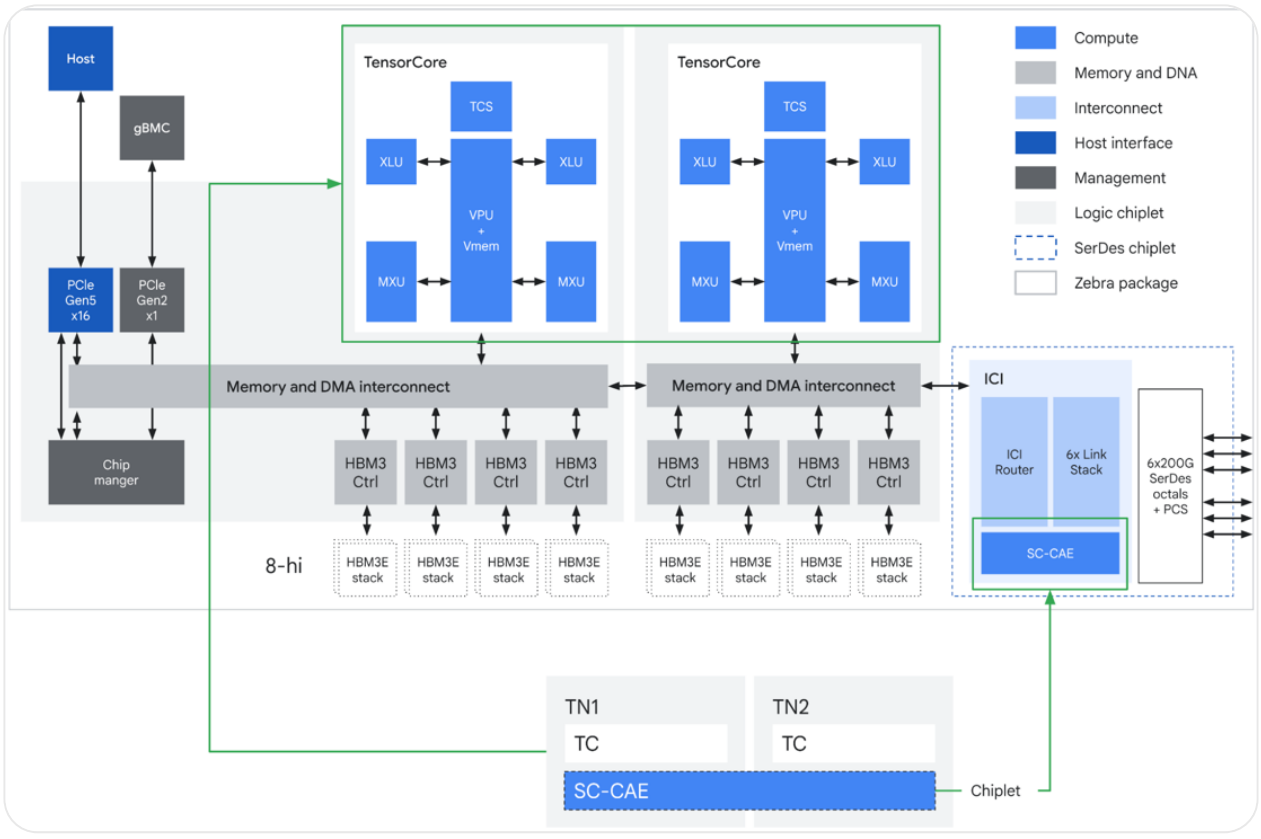


Figure 4: TPU 8i ASIC block diagram

图 4：TPU 8i ASIC 模块框图

- **The Collectives Acceleration Engine (CAE)**: To solve the sampling bottleneck, TPU 8i uses the CAE, which aggregates results across cores with near-zero latency, specifically accelerating the reduction and synchronization steps required during auto-regressive decoding and "chain-of-thought" processing. For each TPU 8i chip, there are two Tensor Cores (TC) on-core dies and one CAE on the chiplet die, replacing four SparseCores (SCs) on core dies in previous-generation Ironwood TPU. By integrating a specialized CAE, TPU 8i further reduces the on-chip latency of collectives by 5x.

集合加速引擎 (CAE)：为了解决采样瓶颈，TPU 8i 采用了 CAE，它能以近乎零的延迟聚合各核心的结果，专门加速自回归解码和“思维链”处理过程中所需的归约与同步步骤。在每个 TPU 8i 芯片中，核心裸片（on-core dies）上有两个张量核心 (TC)，而小芯片裸片（chiplet die）上有一个 CAE，取代了上一代 Ironwood TPU 核心裸片上的四个稀疏核心 (SC)。通过集成专门的 CAE，TPU 8i 将集合通信的片上延迟进一步降低了 5 倍。

Lower latency per collective operation means less time spent waiting, directly contributing to higher throughput required to run millions of agents concurrently.

单次集合操作的延迟降低意味着等待时间减少，这直接有助于提高并发运行数百万个智能体所需的高吞吐量。

- **Boardfly ICI topology**: While the 3D torus allows connecting

thousands of chips to be used in cohesion, a large mesh does have more hops between chips and higher all-to-all latencies. For 8i, we changed how the chips connect together in fully connected boards that are then aggregated into groups.

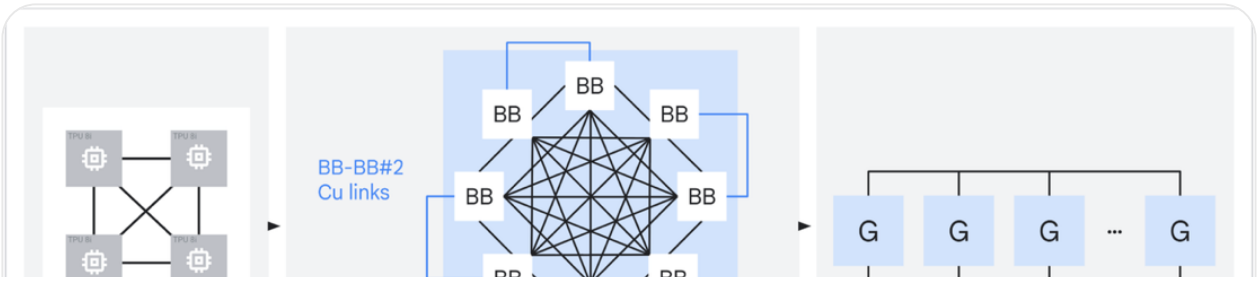
Boardfly ICI 拓扑：虽然 3D 环面（3D torus）允许将数千个芯片连接在一起协同使用，但大型网格在芯片之间确实存在更多跳数，且全对全（all-to-all）延迟较高。对于 8i，我们改变了芯片的连接方式，将其连接在全互连的电路板上，然后再聚合成组。

Utilizing a high-radix design, we connect up-to 1,152 of these chips together, reducing the network diameter and the number of hops a data packet must take to cross the system.

利用高基数（high-radix）设计，我们将多达 1,152 个此类芯片连接在一起，从而减小了网络直径，并减少了数据包跨越系统必须经过的跳数。

By slashing the hops required for all-to-all communication (the heart of MoE and reasoning models), Boardfly achieves up to a 50% improvement in latency for communication-intensive workloads.

通过大幅减少全对全通信（All-to-all communication，MoE 和推理模型的核心）所需的跳数，Boardfly 在通信密集型工作负载中的延迟降低了高达 50%。



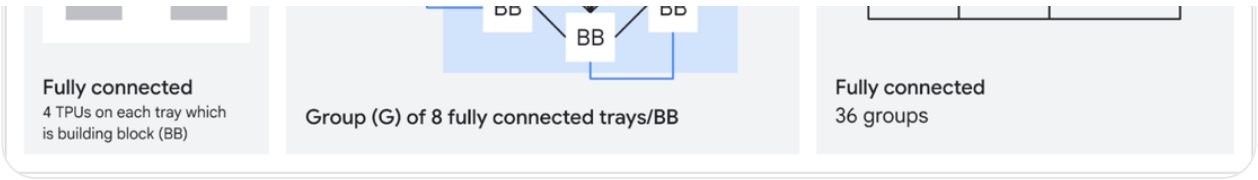


Figure 5: TPU 8i hierarchical Boardfly topology building up from a building block of four fully connected chips into a fully connected group of eight boards, with 36 of such groups fully connected into a TPU 8i pod

图 5：TPU 8i 分层 Boardfly 拓扑结构，从四个全互连芯片组成的构建块开始，扩展到由八块板卡组成的全互连组，其中 36 个此类组全互连形成一个 TPU 8i Pod

Boardfly consists of the following elements, and its topology is hierarchical by nature:

Boardfly 由以下要素组成，其拓扑结构本质上是分层的：

- **Building Block (BB):** Each tray forms a four-chip ring using internal ICI links, providing 16 external connections for broader networking.

  构建块 (BB)：每个托盘利用内部 ICI 链路形成一个四芯片环，并提供 16 个外部连接用于更广泛的网络连接。

- **Group (G):** Eight boards are fully connected via copper cabling to create a localized group, utilizing 11 of the available external links for intra-group communication.

  组 (G)：八块板卡通过铜缆全连接形成一个局部组，利用 11 条可用外部链路进行组内通信。

- **Pod structure:** The final architecture scales to 36 groups (up to 1,024 active chips) linked through Optical Circuit Switches (OCS), ensuring a maximum latency of seven hops for any chip-to-chip

ensuring a maximum latency of seven hops for any chip-to-chip communication.

集群结构：最终架构通过光路交换机 (OCS) 扩展至 36 个组（最多包含 1,024 个活动芯片），确保任何芯片间通信的最大延迟不超过七跳。

Deep dive: The Boardfly vs. torus math

深度解析：Boardfly 与环面 (torus) 拓扑的数学对比

Why move away from the torus for TPU 8i? It comes down to network diameter.

为什么 TPU 8i 不再采用环面拓扑？这归结为网络直径问题。

In a 3D torus, nodes are arranged in a grid where each dimension wraps around like a ring. To reach the furthest possible chip in a 8 x 8 x 16 (1024-chip) configuration, a packet must traverse half the distance of each ring:

在 3D Torus（三维环面）拓扑中，节点排列在网格中，且每个维度都像圆环一样首尾相连。在 8 x 8 x 16（1024 芯片）的配置中，为了到达可能的最远芯片，数据包必须经过每个环的一半距离：

3D torus = 8/2(X) + 8/2(Y) + 16/2(Z) = 16 hops

3D torus = 8/2(X) + 8/2(Y) + 16/2(Z) = 16 跳 (hops)

While the torus is highly efficient for the neighbor-to-neighbor communication typical of dense training, it creates a latency tax for all-to-all communication patterns. In the era of reasoning models and MoE, where any chip may need to talk to any other chip to route a token, this hop count matters.

虽然 Torus 拓扑对于密集训练中典型的邻节点间通信非常高效，但它会为全对全（all-to-all）通信模式带来延迟开销。在推理模型和混合专家模型（MoE）时代，任何芯片都可能需要与其他任何芯片通信以路由 Token，因此这种跳数至关重要。

Boardfly's high-radix topology is inspired by Dragonfly topology principles. By increasing the number of direct optical long-haul links between groups of boards, we flatten the network. For that same 1024-chip pod, Boardfly reduces the network diameter from 16 hops down to just seven.

Boardfly 的高基数（high-radix）拓扑灵感源自 Dragonfly 拓扑原理。通过增加板组之间直接光学长途链路的数量，我们展平了网络。对于同样的 1024 芯片集群（pod），Boardfly 将网络直径从 16 跳减少到了仅 7 跳。

This 56% reduction in network diameter translates directly to lower tail latency, so that the TPU 8i CAE isn't left waiting for data to arrive from across the pod.

网络直径减少了 56%，这直接转化为更低的尾部延迟，从而使 TPU 8i CAE 无需等待数据从整个 Pod 的另一端传输过来。

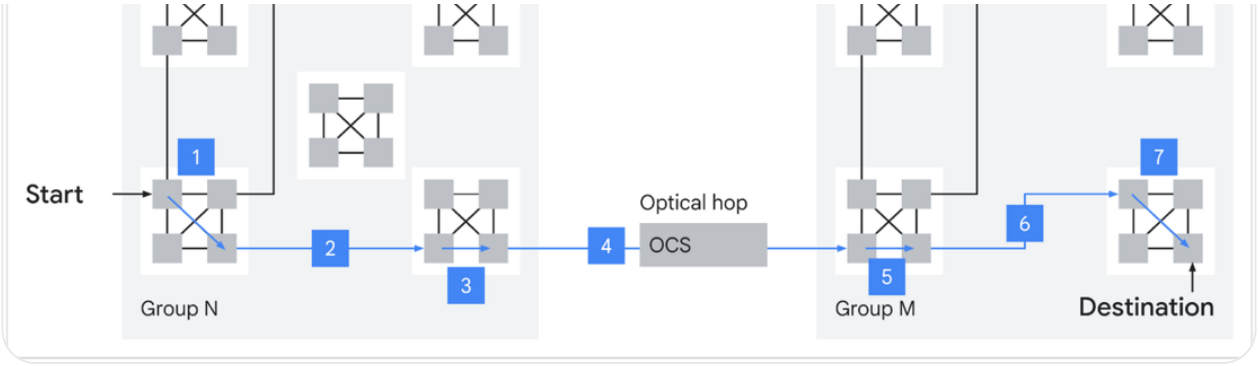


Figure 6: A visual representation of the maximum seven-hop ICI network diameter via optical circuit switch on TPU 8i pod

图 6：TPU 8i Pod 上通过光路交换机实现的七跳最大 ICI 网络直径视觉表示

TPU 8t and TPU 8i at a glance

TPU 8t 与 TPU 8i 一览

Feature	TPU 8t	TPU 8i
Primary Workload 主要工作负载	Large-scale pre-training **大规模预训练**	Sampling, serving, and reasoning **采样、服务与推理**
Network Topology 网络拓扑	3D torus	Boardfly

Specialized Chip Features 专用芯片特性	SparseCore (Embeddings) & LLM Decoder Engine **SparseCore (嵌入) 与 LLM 解码引擎**	CAE (Collectives Acceleration Engine) **CAE (集合通信加速引擎)**
HBM Capacity	216 GB	288 GB
On-Chip SRAM (Vmem) 片上 SRAM (Vmem)	128 MB	384 MB
Peak FP4 PFLOPs FP4 算力峰值 (PFLOPs)	12.6	10.1
HBM Bandwidth	6,528 GB/s	8,601 GB/s (~1.3x of TPU 8t) **8,601 GB/s （约为 TPU 8t 的 1.3 倍）**
CPU Header	**Arm Axion**	**Arm Axion**

Software enablement: A performance-first AI

stack

软件赋能：性能优先的 AI 技术栈

Hardware is only as powerful as the software that drives it. The eighth generation of TPUs are built on the same performance-first stack we pioneered with the seventh-generation Ironwood TPUs, designed to make custom kernel development accessible without sacrificing the abstraction of high-level frameworks. This stack includes:

硬件的强大程度取决于驱动它的软件。第八代 TPU 构建在我们为第七代 Ironwood TPU 开创的性能优先技术栈之上，旨在让自定义内核开发变得触手可及，同时又不牺牲高级框架的抽象性。该技术栈包括：

- **Pallas and Mosaic**: We provide first-class support for Pallas, our custom kernel language that lets you write hardware-aware kernels in Python. This enables you to squeeze every drop of performance out of the TPU 8i CAE and the TPU 8t SparseCore.

  Pallas 和 Mosaic：我们为 Pallas 提供一流的支持，这是我们的自定义内核语言，允许您使用 Python 编写硬件感知内核。这使您能够榨干 TPU 8i CAE 和 TPU 8t SparseCore 的每一滴性能。

- **Native PyTorch experience:** We're thrilled to share that native PyTorch support for TPUs is now in preview. If you're currently building and serving models on PyTorch, we've made it easier than ever to start using TPUs. You can bring your existing models to our

over to start using TPUs. You can bring your existing models to our TPUs just as they are, complete with full support for the native features you rely on, such as Eager Mode.

原生 PyTorch 体验：我们非常激动地宣布，TPU 对原生 PyTorch 的支持现已进入预览阶段。如果您目前正在使用 PyTorch 构建和提供模型服务，那么现在开始使用 TPU 将比以往任何时候都更加容易。您可以直接将现有模型迁移到我们的 TPU 上，并获得对您所依赖的原生功能（如 Eager Mode）的完整支持。

- **Portability**: The same JAX, PyTorch, or Keras code that runs on Ironwood scales to this generation. Accelerated Linear Algebra (XLA) handles the complex translation of the Broadly topology and CAE synchronization behind the scenes, allowing you to focus on your model, not the interconnect.

可移植性：在 Ironwood 上运行的 JAX、PyTorch 或 Keras 代码同样可以扩展到这一代产品。加速线性代数 (XLA) 在后台处理 Broadly 拓扑和 CAE 同步的复杂转换，让您可以专注于模型本身，而非互连技术。

Generation over generation: The performance leap

代际更迭：性能飞跃

Our commitment to co-designing hardware and software continues to pay dividends. When compared to seventh-generation Ironwood TPU, the eighthgeneration TPUs deliver massive gains:

我们对硬件和软件协同设计的承诺持续带来回报。与第七代

Ironwood TPU 相比，第八代 TPU 实现了巨大的提升：

- **Training price-performance**: TPU 8t delivers up to 2.7x performance-per-dollar improvement over Ironwood TPU for large-scale training.

  训练性价比：在进行大规模训练时，TPU 8t 相比 Ironwood TPU 实现了高达 2.7 倍的单位成本性能提升。

- **Inference price-performance**: TPU 8i delivers up to 80% performance-per-dollar improvement over Ironwood TPU, particularly at low-latency targets for large MoE models.

  推理性价比：TPU 8i 相比 Ironwood TPU 实现了高达 80% 的单位成本性能提升，在大型 MoE 模型的低延迟目标下表现尤为出色。

- **Energy efficiency**: Both chips deliver up to 2x better performance-per-watt, critical for scaling the next generation of AI sustainably.

  能效比：两款芯片均实现了高达 2 倍的单位功耗性能提升，这对于可持续地扩展下一代 AI 至关重要。

Looking ahead

To empower Google Cloud customers pioneering the next wave of innovation, we designed TPU 8t and TPU 8i as two distinct, specialized systems tailored to the multifaceted future demands of the AI lifecycle.

为了助力引领下一波创新浪潮的 Google Cloud 客户，我们将

TPU 8t 和 TPU 8i 设计为两个独立且专业的系统，以量身定制的方式满足 AI 生命周期中多元化的未来需求。

TPU 8t and 8i are both purpose-built for the most demanding serving and training workloads, fully integrating with the AI Hypercomputer software stack: JAX, PyTorch, vLLM, XLA, and Pathways. This specialization and ground-up redesign, all in deep collaboration with Google Deepmind, delivers exceptional price-performance and power efficiency.

TPU 8t 和 8i 均专为最严苛的推理和训练工作负载而打造，并与 AI Hypercomputer 软件栈（JAX、PyTorch、vLLM、XLA 和 Pathways）深度集成。这种专业化设计以及与 Google Deepmind 深度合作进行的底层重构，提供了卓越的性价比和能效比。

The modularity of our eighth-generation architecture provides a clear and unique roadmap for the future. Just as every major shift in the computing landscape has required infrastructure breakthroughs, so does the agentic era.

我们第八代架构的模块化特性为未来提供了清晰且独特的路线图。正如计算领域的每一次重大变革都需要基础设施的突破一样，智能体（Agentic）时代也不例外。

Reasoning agents that plan, execute, and learn within continuous feedback loops cannot operate at peak efficiency on hardware that was originally optimized for traditional training or transactional inference;

their operational intensity are fundamentally distinct.

在持续反馈循环中进行规划、执行和学习的推理智能体，无法在最初为传统训练或事务性推理优化的硬件上发挥最高效率；它们的运行强度在本质上是截然不同的。

Our eighth-generation TPU infrastructure has evolved to meet these specific requirements head-on.

我们的第八代 TPU 基础设施已经过演进，旨在正面满足这些特定需求。

To learn more about the eighth-generation TPU family:

了解更多关于第八代 TPU 系列的信息：

- Submit an interest form for eighth-generation TPUs

  提交第八代 TPU 意向表

- Get involved in the community forums

  参与社区论坛

- Check out the eighth-generation TPU announcement video

  观看第八代 TPU 发布视频

- Visit our TPU website **访问我们的 TPU 网站**
