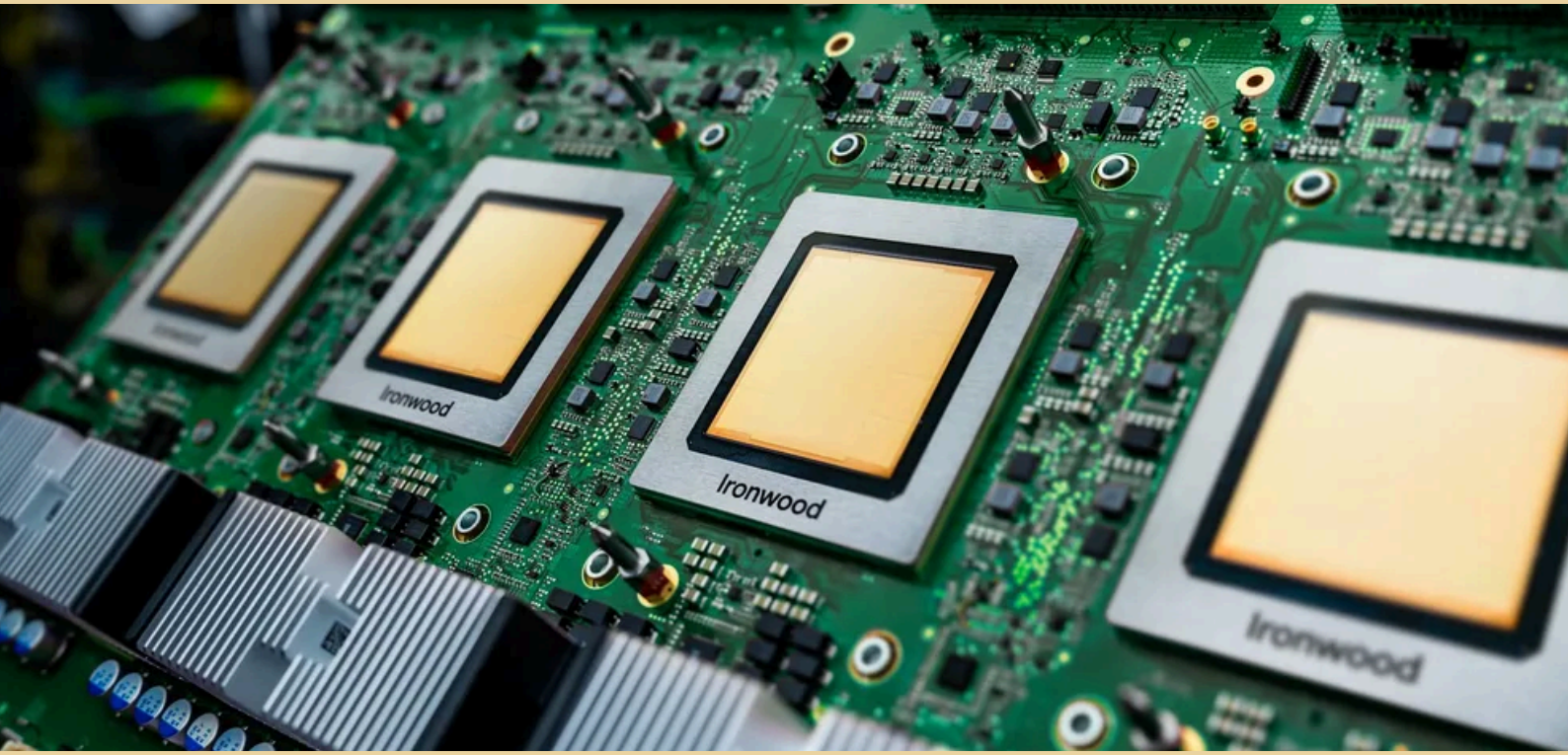
Google's TPU and the New Economics of AI Deployment



ARPU Intelligence

Dec 2025

Executive Summary

The semiconductor industry is entering a new phase of competition. For the last decade, the market for AI compute has been a monopoly defined by a single metric: training performance. In this era, Nvidia was the undisputed victor.

However, the market is now shifting from Training (research) to Inference (production). As AI models move from the lab to high-volume commercial deployment, the primary constraint shifts from "speed at any cost" to "performance per dollar."

This report analyzes how Google is positioning its Tensor Processing Unit (TPU) to capture this new market. Our analysis identifies three structural shifts that challenge Nvidia's dominance:

The Economic Pivot to Inference

While frontier training drives new CapEx, inference now accounts for an estimated 80-90% of total AI compute power usage. As this volume dominance translates into cost dominance, the "Nvidia Tax"—the premium paid for general-purpose GPUs—becomes unsustainable for production fleets. Google is exploiting this by positioning the TPU as the specialized engine for the high-volume inference era.

The Systolic Efficiency Advantage

The TPU's cost advantage is physical. By utilizing a Systolic Array architecture and novel SparseCores for embeddings, Google achieves radically higher efficiency. Independent analysis suggests the TPU v7 (Ironwood) offers an all-in TCO that is ~40% lower than comparable Nvidia GB200 systems, a margin that translates directly to profitability for AI operators.

The Merchant Silicon Strategy

Google has fundamentally altered its TPU strategy. It has moved from using TPUs solely as an internal tool to acting as a merchant vendor. The definitive signal is the Anthropic partnership: a deal involving 1 million TPUs (including 400k direct sales) to power the training of Anthropic's Claude 4.5 Opus. This proves the TPU ecosystem is now mature enough for the world's most demanding frontier training workloads, not just inference.

The Bottom Line:

The future of AI infrastructure is not a monoculture. While Nvidia retains its fortress, Google has successfully opened a second front. For tech operators, the TPU represents the most viable path to escaping hardware lock-in, optimizing unit economics, and gaining critical leverage in the AI supply chain.

Key Google Cloud TPU Customers	
Company	Description
Anthropic	A major strategic partner that uses Google Cloud TPUs for both AI model training and inference. The relationship is a deep collaboration involving commitments for massive compute capacity, reportedly up to a million TPUs.
Apple	Uses Google Cloud TPUs through a business-to-business cloud service agreement to train its internal AI models (Apple Intelligence).
Meta	Currently in discussions about a potential multi-billion dollar deal to purchase and deploy Google TPUs in its own data centers starting in 2027, and to rent TPU capacity from Google Cloud as early as 2026. This signifies a potential shift toward a direct hardware supply relationship alongside cloud services.
Midjourney	A key Google Cloud customer, using TPUs as a cloud service to power its AI-driven image generation workloads.
Safe Superintelligence (SSI)	An AI startup that accesses Google TPUs via Google Cloud services for its development needs.
Salesforce	Uses Google TPUs through Google Cloud for its AI and machine learning initiatives within its enterprise software offerings.
Internal Google Use	Google itself is the primary user of TPUs, deploying them across its various services (Search, Gmail, Maps, Gemini AI models, etc.) to power AI features for billions of users. This is an internal, integrated relationship.

Table of Contents

The Pivot to Inference

Why AI Workloads Are Shifting from Training to Inference

For the past decade, the AI hardware market has been driven by a single imperative: **Discovery**. The goal was to build larger, smarter models at any cost. In this regime, the primary metric was "Time-to-Train," and Nvidia's GPUs were the only logical choice. Cost was secondary to capability.

In 2025, the industry has entered a new phase: **Utility**. The foundational models (GPT-4, Gemini, Claude) are increasingly intelligent. Now, the imperative is to integrate them into every product in the economy. This transition fundamentally alters the economic physics of the data center.

Why Inference Favors the Specialist

The shift to inference creates a specific technical opening for the TPU because the workload characteristics change fundamentally:

- **Training is Dynamic:** Training a model is an experimental process. Algorithms change weekly (e.g., FlashAttention, Mixture of Experts). Researchers need a programmable, flexible chip (the GPU) that can adapt to new code instantly. Nvidia's CUDA ecosystem ensures that every new research breakthrough runs on their hardware on Day 1. This ecosystem velocity is why Nvidia retains its grip on training—switching to a TPU slows down the research cycle.
- **Inference is Static:** Once a model is trained, its weights are frozen. The computational graph becomes fixed and predictable. This removes the need for extreme flexibility. You no longer need a chip that can do *anything*; you need a chip that can execute *this specific graph* as efficiently as possible.

This stability allows operators to port their models to specialized hardware (ASICs) like the TPU. The engineering cost of migration is paid once, but the efficiency dividends are reaped on every single query.

The Volume/Value Inversion

The market is currently defined by a divergence between capital expenditure and operational reality.

- **The CapEx Reality:** The majority of *new* hardware spending is still focused on training clusters. Hyperscale CapEx is growing at over 50% year-over-year, driven by the purchase of massive Nvidia H100 and Blackwell clusters.

- **The Operational Reality:** The actual *work* being done has already flipped. According to MIT Technology Review, inference now accounts for an estimated 80-90% of total AI compute power usage**.**

This creates a Volume/Value Inversion. While the headline checks are written for training, the daily operational cost is overwhelmingly driven by inference.

The Inference Surge

As AI moves from R&D to production, operators face an Inference Surge. Industry data indicates that inference costs can balloon by 500-1,000% during this scaling phase. A $30,000 GPU that makes sense for a high-value training run becomes economically ruinous when used to serve millions of low-margin queries.

For a tech operator, the math is simple: Training is a one-time CapEx investment; Inference is a permanent OpEx tax.

The Leverage Effect

The mere existence of the TPU is already altering market pricing. Reports indicate that OpenAI saved ~30% on its Nvidia fleet costs simply by leveraging the competitive threat of adopting TPUs.

This reveals the value of the TPU for operators: even if you don't migrate 100% of your workload, maintaining a credible TPU pilot creates immense negotiating leverage against the "Nvidia Tax."

The Nvidia Tax in Production

Nvidia GPUs are general-purpose supercomputers. They carry silicon "baggage" designed for graphics and scientific simulation that is unnecessary for repetitive inference. Using them for high-volume serving is like using a Ferrari to deliver pizza.

To survive the Inference Surge, operators need a delivery van. They need a chip that sheds flexibility for ruthless efficiency. This is the market gap Google’s TPU fills, offering an estimated 40% lower TCO for comparable workloads.

The Physics of Efficiency

Why the TPU Wins on Unit Economics

To understand Google's economic advantage, you have to look at the silicon itself. The battle between the TPU and the GPU is a clash of two fundamentally different architectural philosophies.

Nvidia's GPUs are built on a **SIMT (Single Instruction, Multiple Threads)** architecture. This design is versatile but carries significant silicon "baggage" to support general-purpose computing. Google's TPU sheds this entirely, using an ASIC design optimized for one thing: the matrix math of AI.

The Systolic Array Advantage

The core of the TPU is the **Systolic Array**. Unlike a GPU, which constantly moves data back and forth between memory and cores, a systolic array pumps data through a grid of processors in a rhythmic wave.

- **Data Reuse:** Data is read from memory once and passed directly between processing units, drastically reducing energy consumption.
- **SparseCores:** To address a historical weakness in handling non-dense data (like recommendation embeddings), modern TPUs integrate specialized "SparseCores." These embedded units handle data-scatter/gather operations, allowing the main array to stay focused on math, ensuring high performance even for complex, irregular workloads.

The Result: Studies show recent TPUs delivering 1.7–3x greater performance per watt than comparable GPUs. In a power-constrained data center, this efficiency is the primary driver of margin.

Optical Circuit Switching (OCS): The Scale Advantage

Google's most important advantage may be the wires that connect the chips. Instead of using expensive, power-hungry electrical switches (InfiniBand), Google uses Optical Circuit Switching (OCS) to reflect beams of light between racks.

- **Massive World Size:** This optical fabric allows Google to connect up to 9,216 TPU v7 (Ironwood) chips into a single, unified supercomputer pod. This "world size" significantly exceeds standard GPU clusters, reducing the latency penalties that usually plague massive training runs.

- **Dynamic Reconfiguration:** If a chip fails, OCS can physically reroute the light path in milliseconds to bypass the node, maintaining high uptime for weeks-long training jobs.

The TCO Equation: Bypassing the Nvidia Tax

When you combine the silicon efficiency with the optical fabric, the cost advantage is staggering.

Analysts estimate that Google's vertical integration allows it to operate its TPU infrastructure at an all-in TCO that is ~40% lower than a comparable Nvidia GB200 server cluster.

For the customer, this structural advantage translates to aggressive pricing—often 2x lower per hour than equivalent Nvidia instances—while Google retains healthy cloud margins. Simply put, the TPU delivers more intelligence for every dollar.

Unit Economics Comparison (Nvidia Blackwell vs. Google TPU v7)

Company	Nvidia GB200 NVL72 (Hyperscale Cost)	Nvidia GB300 NVL72 (Hyperscale Cost)	Google TPU v7 (Internal Cost)	Google TPU v7 (External Price)
Total Cost per Chip/Hour	$2.28	$2.73	$1.28	$1.60
Capital Cost %	77.4%	79.0%	72.7%	72.7%
Memory Capacity	192 GB	288 GB	192 GB	192 GB
TCO per FP8 PFLOP	$0.46	$0.55	$0.28	$0.40

Source: Semianalysis

Note:
- "External Price" reflects estimated pricing for high-volume customers like Anthropic, inclusive of Google's margin.
- Hourly TCO estimates assume a 5-year useful economic life for the hardware. 'Capital Cost %' represents the portion of the hourly rate attributed to hardware amortization (chip, server, and networking CapEx) versus operating expenses (electricity and cooling).
- TCO per FP8 PFLOP measures exactly how much it costs to buy one unit of AI capability (one PetaFLOP of compute power) for one hour.

The Software Bridge

How Google Is Bridging the Software Gap

Software compatibility has long been the primary barrier to adopting hardware alternatives to Nvidia. Most AI applications are built on CUDA, Nvidia's proprietary platform, which creates high switching costs for engineering teams.

In 2025, Google's strategy is designed to directly address this challenge. By focusing on compatibility with industry-standard frameworks, Google is creating a viable software path for developers to migrate workloads from GPUs to its more cost-effective TPU infrastructure.

PyTorch on TPU: The "No-Retraining" Playbook

Google's most important strategic move is its full-throated support for **PyTorch**, the industry's most popular AI framework.

- **The Problem:** Previously, running PyTorch on a TPU was a clunky, inefficient process that frustrated developers and caused major partners like Meta to abandon the platform.
- **The 2025 Fix:** Google is re-architecting its software to support a native PyTorch backend.
- **The Impact:** This move is a game-changer for enterprise adoption. It allows a company's engineering team to take their existing PyTorch models—the ones already built for Nvidia GPUs—and deploy them on more cost-effective TPU infrastructure with minimal code changes. It dramatically reduces the financial and operational risk of migration, turning it from a multi-year re-architecture project into a simple deployment decision.

JAX: The High-Performance Option

While PyTorch offers compatibility, JAX offers peak performance.

- **The Analogy:** Think of JAX as the "manual transmission" for AI. It's harder to learn than the "automatic" of PyTorch, but for expert teams who need to squeeze every last drop of performance out of their hardware, it offers unparalleled control.
- **The Validation:** The world's most advanced AI labs—including those building Gemini 3, Grok (xAI), and Claude—use JAX. This proves that for the most demanding workloads on Earth, the TPU software stack is not just a viable alternative; it is the preferred choice.

vLLM & SGLang: The Open Inference Standard

The final piece of the bridge is the serving layer. Most enterprises use open-source standards like vLLM to run their models in production because they are highly optimized for cost and speed.

- **The Integration:** Google has dedicated a massive engineering effort to integrate TPU support directly into these open-source projects.
- **The Impact:** This is a crucial move for commoditizing the inference stack. It means an operator can use the same open-source tools to deploy a model on a TPU as they would on a GPU. This removes vendor lock-in at the serving layer, enabling companies to route inference jobs to the most cost-effective hardware in real time based on price and availability, rather than being locked into a single vendor's proprietary serving solution.

By embracing PyTorch, perfecting JAX, and contributing to open standards, Google is systematically dismantling the software barriers that once protected Nvidia's monopoly.

Beyond the Chip

The Optical Circuit Switching Advantage

The economic and silicon advantages of the TPU establish it as a formidable competitor to Nvidia's GPUs. However, the TPU's true, long-term defensible advantage is not found on the die, but in the architecture of the data center that surrounds it. Google's decade of experience operating at planetary scale has produced a system-level architecture—spanning networking, cooling, and resilience—that is fundamentally different from the merchant market and difficult, if not impossible, for competitors to replicate.

Optical Circuit Switching (OCS)
While Nvidia has masterfully solved chip-to-chip communication within a rack (NVLink), Google’s critical innovation is solving the far harder problem of rack-to-rack communication at an unprecedented scale. The key is its unique Optical Circuit Switching (OCS) interconnect.

Unlike a standard electrical network switch (like InfiniBand) which consumes significant power converting optical signals to electrical and back again, OCS physically reflects beams of light. This approach yields profound architectural benefits:

- **Unprecedented "World Size":** OCS is the technology that enables a single cluster to scale to 9,216 Ironwood TPUs with a combined 1.77 PB of HBM memory. As *The Next Platform* put it, this scale "makes a rackscale Nvidia system...look like a joke." This isn't just a larger number; it allows for training runs on frontier models and massive inference workloads to operate as a single, cohesive unit, minimizing the cross-cluster communication penalties that plague traditional designs.
- **Architectural Tradeoffs:** The topology enabled by OCS is a 3D Torus, which is exceptionally efficient for workloads with predictable, nearest-neighbor communication patterns. This contrasts with Nvidia's switched NVLink fabric, which is closer to a "fat tree" topology. Nvidia's approach can be superior for workloads requiring frequent, chaotic all-to-all communication (like some Mixture of Experts models). However, Google's design represents a deliberate, cost-effective tradeoff optimized for the majority of large-scale AI tasks, betting that extreme scalability and efficiency trump universal flexibility.

Resilience and Real-World Goodput

A theoretical performance advantage is meaningless if the system is down. For multi-week training jobs, a single node failure can be catastrophic, forcing a costly restart. This is where the true operational value of Google's system design becomes clear.

The software-defined nature of the OCS fabric allows for dynamic reconfiguration. If a TPU or a link fails, the control plane can "heal" the network in milliseconds by physically rerouting the light paths around the fault. This capability to maintain the integrity of a large cluster for the entire duration of a job is a critical, underappreciated advantage. It transforms the system from a fragile collection of nodes into a resilient, self-healing fabric, maximizing "goodput"—the actual, useful work completed—rather than just peak theoretical FLOPs. This is further complemented by Google's mature, fifth-generation liquid cooling infrastructure, which ensures thermal stability and reliability across these massive fleets.

Selling an Architectural Philosophy

This system-level mastery is a moat born from necessity—the necessity of serving Google's own immense internal workloads. Now, with the shift to a merchant strategy, More than selling a chip, Google is selling its entire, battle-tested hyperscale architecture as a product.

When Anthropic buys or leases a million TPUs, it is not just acquiring silicon. It is acquiring access to a pre-packaged, vertically integrated supercomputing solution that includes the optical fabric, the cooling systems, and the control plane software. This is a fundamentally different value proposition than buying GPUs and networking components off the shelf and attempting to integrate them. Google is leveraging a decade-long, in-house R&D advantage in system design and offering it directly to the market, a feat that no other cloud provider or hardware vendor can currently match.

Implications For Your Role

Major market events create cascading effects that ripple across the entire ecosystem, impacting every business and function in a unique way.

Google's decision to become a "merchant silicon" vendor for TPUs—selling and leasing them directly to hyperscalers and enterprises—is not just a new product launch; it's the first domino in a fundamental restructuring of the AI supply chain. This move creates new leverage for buyers, new risks for incumbents, and a new strategic calculus for anyone building or buying AI infrastructure at scale.

While most analysis focuses on the initial event, ARPU specializes in mapping these cascading effects directly to your specific role and business model.

To demonstrate, let's deconstruct the single most important strategic challenge from this report—The End of the Nvidia Monopoly—and show how it creates distinct, actionable intelligence for different tech operators.

One Trend, Three Bespoke Implications: Navigating a Duopoly

The emergence of the TPU as a viable, at-scale alternative to Nvidia GPUs creates a new set of strategic imperatives for operators across the stack.

⚙️ **For the Head of Infrastructure / CTO at a Large Enterprise or AI Startup:**

You now have negotiating leverage. For the last five years, your AI roadmap has been dictated by Nvidia's pricing and allocation. The validation of the TPU by Anthropic and Apple means you can now build a credible, multi-vendor hardware strategy. A pilot program on TPUs is no longer just a technical experiment; it is a powerful procurement tool. The threat of moving even 20% of your inference workload to TPUs can unlock significant discounts (reports suggest up to 30%) from your incumbent GPU provider.

The Key Question: *How are you leveraging a multi-vendor cloud and hardware strategy not just for technical resilience, but as an active tool to reduce your multi-million dollar AI infrastructure bill?*

🎯 **For the Head of Product at an AI Infrastructure Software Company (e.g., MLOps, Data Platforms):**

Your product roadmap must now account for a heterogeneous compute environment. Building exclusively for the CUDA ecosystem is no longer a viable long-term strategy. The Anthropic deal proves that your largest potential customers will have significant, non-Nvidia infrastructure. If your software does not support PyTorch/XLA and is not optimized for TPUs, you risk being locked out of a massive and growing segment of the market.

The Key Question: *Does your 2026 product roadmap include a dedicated engineering effort to ensure your platform is performant and certified on Google's TPU infrastructure, or are you at risk of becoming a "CUDA-only" legacy tool?*

🤝 **For the Head of Strategy at a Competing Cloud Provider or Neocloud:**

Google has just weaponized its balance sheet to finance its partners' growth. The "Hyperscaler Backstop" model used to fund Anthropic's data center expansion is a new competitive threat. Google is effectively offering off-balance-sheet financing to lock in strategic customers and deny capacity to Nvidia. To compete, you can no longer just offer the best hardware; you must now be prepared to offer creative financing solutions that match this new go-to-market motion.

The Key Question: *What is your capital strategy to compete with a rival who is willing to underwrite their customers' infrastructure risk, and which financial partners can you bring to a deal to level the playing field?*

Conclusion

The New Economics of AI at Scale

The AI hardware market has reached an inflection point. For the last decade, Nvidia's dominance was absolute because the market was defined by a single constraint: the difficulty of training. Today, as AI permeates the global economy, the constraints are shifting to power, cost, and the logistics of **planetary-scale deployment.**

Google's TPU strategy is no longer a defensive hedge; it is an offensive disruption. It began by attacking the Inference Surge with a vertically integrated, cost-optimized architecture. It is now systematically dismantling the software barriers that protected Nvidia's monopoly. But the definitive, and perhaps most durable, advantage is not in the silicon alone—it is in the system that surrounds it. Through its unique OCS fabric, Google has solved the challenge of rack-to-rack communication at a scale and resilience that competitors cannot currently match.

The evidence for this new competitive reality is clear and builds upon four distinct layers:

- **The Physics:** At the core, the TPU's Systolic Array delivers fundamentally better performance per watt for the matrix math that defines AI.
- **The Economics:** Vertical integration and a specialized design bypass the "Nvidia Tax," offering a structural TCO advantage of over 40% for comparable workloads when it comes to AI inference.
- **The Bridge:** Pragmatic support for native PyTorch and open-source standards like vLLM is systematically lowering switching costs and eroding the CUDA moat.
- **The Scale:** The OCS interconnect provides a defensible system-level moat, enabling massive "world size" and operational resilience that transforms the economics of deploying frontier models.

For the tech operator, this means choice. The future of the data center is not a monoculture of general-purpose GPUs. It is a heterogeneous fleet, where a model's deployment will be dictated not just by its architecture, but by the scale and unit economics at which it must operate.

About ARPU

ARPU provides bespoke, function-specific intelligence for tech operators. In a market saturated with generic research and noise, we deliver signal.

Our analysis translates market-wide events and trends into intelligence briefs tailored to the specific challenges and questions of your role. We exist to help you make better, faster decisions.

Go Deeper.

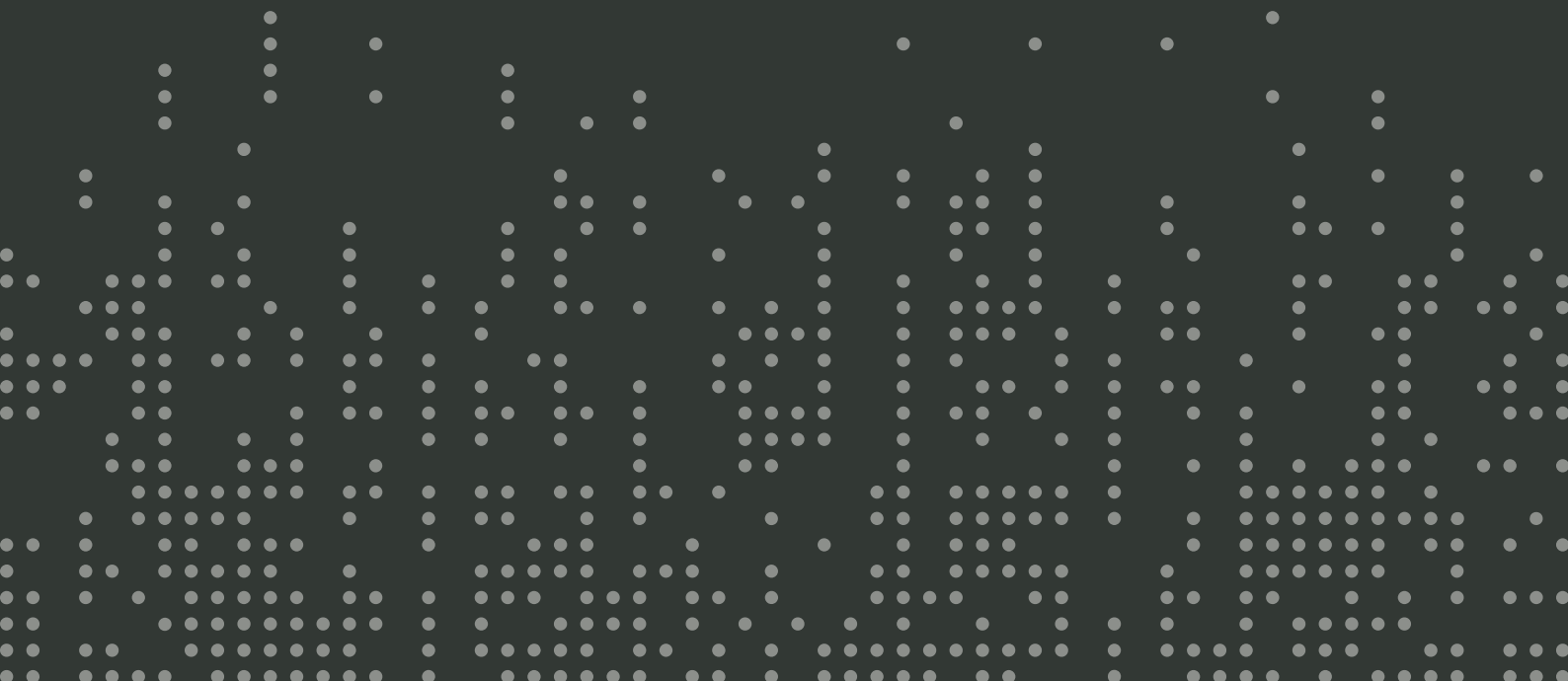
This report provided a market-wide view of the TPU landscape. But your organization, your products, and your strategic questions are unique.

Ready to see what these trends mean for your 2026 roadmap?

Request a Complimentary, Personalized Sample Brief

Contact us directly at arpu@hedder.com

Or visit our website at https://arpu.hedder.com/



Sources

Apple Trained its Apple Intelligence Models on Google TPUs, Not NVIDIA GPUs https://www.techpowerup.com/325066/apple-trained-its-apple-intelligence-models-on-google-tpus-not-nvidia-gpus

Awesome OpenXLA https://openxla.org/stablehlo/awesome

Google Shows Off Its Inference Scale And Prowess https://www.nextplatform.com/2025/09/17/google-shows-off-its-inference-scale-and-prowess/

Google TPU v6e vs GPU: 4x Better AI Performance Per Dollar Guide https://introl.com/blog/google-tpu-v6e-vs-gpu-4x-better-ai-performance-per-dollar-guide

GPU and TPU Comparative Analysis Report https://bytebridge.medium.com/gpu-and-tpu-comparative-analysis-report-a5268e4f0d2a

Hugging Face and Google partner for open AI collaboration https://huggingface.co/blog/gcp-partnership

Hyperscaler Blackwell and Custom Accelerator Rollouts Drive 53 Percent Capex Growth in 1Q 2025 https://www.delloro.com/news/hyperscaler-blackwell-and-custom-accelerator-rollouts-drive-53-percent-capex-growth-in-1q-2025/

JAX on GPUs: Implementation Strategies for Enterprise Machine Learning https://lambda.ai/blog/pytorch-to-jax-on-lambda-for-enterprise-ml

Midjourney Selects Google Cloud to Power AI-Generated Creative Platform https://www.googlecloudpresscorner.com/2023-03-14-Midjourney-Selects-Google-Cloud-to-Power-AI-Generated-Creative-Platform

Open source collaborations and key partnerships to help accelerate AI innovation https://cloud.google.com/blog/products/ai-machine-learning/googles-open-source-momentum-openxla-new-partnerships

PyTorch/XLA 2.5: vLLM support and an improved developer experience https://cloud.google.com/blog/products/ai-machine-learning/whats-new-with-pytorchxla-2-5

The chip made for the AI inference era – the Google TPU https://substack.com/home/post/p-179815720

The Cost of AI Compute: Google's TPU Advantage vs. OpenAI's Nvidia Tax https://www.nasdaq.com/articles/cost-ai-compute-googles-tpu-advantage-vs-openais-nvidia-tax

The cost of compute: A $7 trillion race to scale data centers https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/the-cost-of-compute-a-7-trillion-dollar-race-to-scale-data-centers

TPU architecture https://docs.cloud.google.com/tpu/docs/system-architecture-tpu-vm

TPU vs GPU: Comprehensive Technical Comparison https://www.wevolver.com/article/tpu-vs-gpu-in-ai-a-comprehensive-guide-to-their-roles-and-impact-on-artificial-intelligence

TPUv7: Google Takes A Swing At The King https://newsletter.semianalysis.com/p/tpuv7-google-takes-a-swing-at-the

We did the math on AI's energy footprint. Here's the story you haven't heard. https://www.technologyreview.com/2025/05/20/1116327/ai-energy-usage-climate-footprint-big-tech/

What's new with Google Cloud's AI Hypercomputer architecture https://cloud.google.com/blog/products/compute/whats-new-with-google-clouds-ai-hypercomputer-architecture