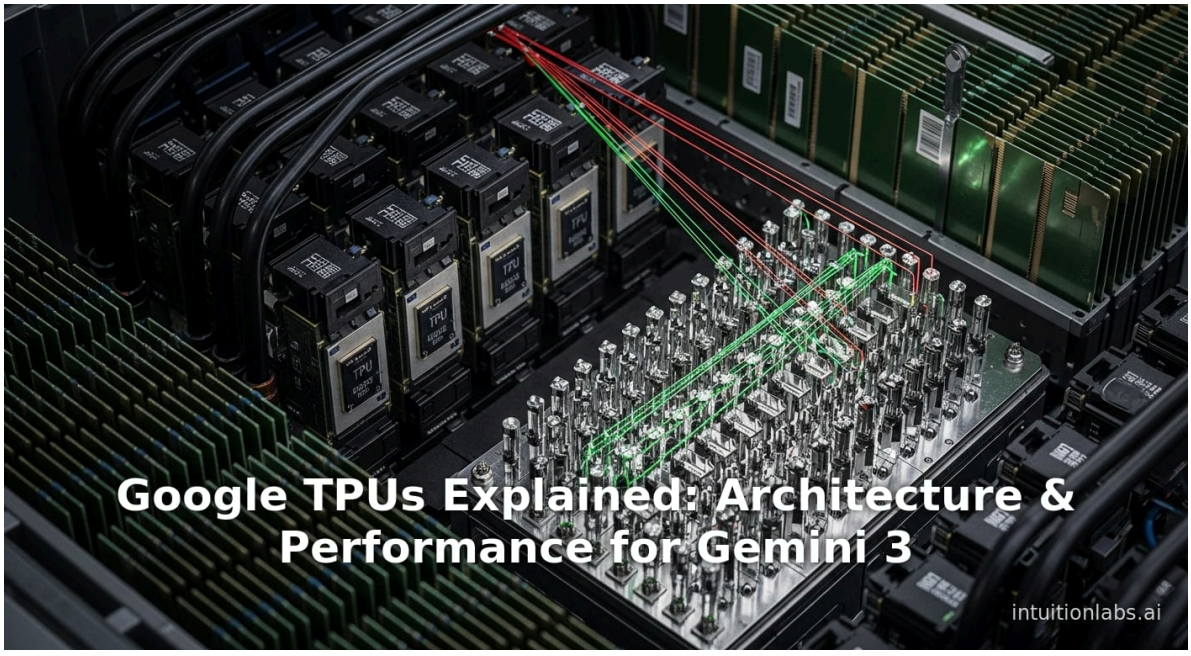
Google TPUs Explained: Architecture & Performance for Gemini 3

By Adrien Laurent, CEO at IntuitionLabs • 12/19/2025 • 35 min read

google tpu | gemini 3 | tpu architecture | ai hardware | tpu vs gpu | ai accelerator | tensor processing unit | machine learning hardware



Executive Summary

The Google Tensor Processing Unit (TPU) is a custom domain-specific accelerator that Google has developed and iterated on since 2015 to handle the exploding demands of machine learning, especially neural network training and inference. Over successive generations (TPU v1 through v7), Google has consistently boosted raw throughput, memory bandwidth, and energy efficiency while specializing its hardware for the "tensor" operations at the heart of modern AI models. TPU improvements include giant systolic arrays (e.g. the v1 TPU had a 256×256 matrix unit with 65,536 MACs and 92 TOPS peak ([1] www.researchgate.net)), high-bandwidth on-chip memory (e.g. v3 boards with 128 GB of HBM yielding over 100 petaflops per 32 TB pod ([2] www.slideshare.net) ([3] www.slideshare.net)), and novel interconnects (optical circuit switching in TPU v4 ([4] arxiv.org)) to scale to thousands of chips. Each new TPU generation has dramatically outpaced prior NVIDIA GPU solutions: for example, TPU v4 pods deliver 1.2–1.7× higher throughput than comparable Nvidia A100 setups while using only 53–77% of the power ([5] arxiv.org).

These TPU advances have made them the "powerhouse behind" Google's latest large language model, Gemini 3. According to Google, Gemini 3's training and inference rely entirely on TPU hardware (the v5e and v6e pods) rather than third-party GPUs ([6] www.linkedin.com) ([7] www.itpro.com). By co-designing the chip, compiler, and software stack (Pathways, XLA, JAX) in tandem, Google has achieved substantial cost and performance gains: internal analysis shows TPU pods have lower training cost-per-FLOP than GPU-based clusters and continual per-generation boosts in performance-per-watt ([8] www.linkedin.com). Real-world customers attest to the benefits: AssemblyAI reports up to **4×** better throughput-per-dollar on TPU v5e for speech AI, and Gridspace sees **5×** training speedups on TPUs ([9] cloud.google.com) ([10] cloud.google.com).

This report provides a comprehensive technical analysis of Google's TPU architecture and ecosystem, especially in light of its role in enabling Gemini 3. We review the design of each TPU generation, dataflow and interconnect innovations, integration with Google's software stack, and performance/cost comparisons with GPUs. We highlight case studies and benchmarks, discuss power and efficiency metrics (e.g. TPU v4 pods use **~3×** less energy and emit **~20×** lower CO_2 than comparable on-premise GPU systems ([5] arxiv.org)), and examine the strategic implications for the AI hardware landscape. We then explore future directions, including TPU v7 ("Ironwood") and beyond, and how this bespoke approach to AI chips may reshape large-scale model training and inference. Throughout, we substantiate all claims with extensive citations to academic papers, industry whitepapers, and expert analyses.

Introduction and Background

As deep learning workloads have grown from sensory perception to large-scale language and multimodal models, the demands on hardware have skyrocketed. By the mid-2010s, researchers recognized that "major improvements in cost-energy-performance" for neural network workloads would require *domain-specific hardware*, not just scaling general-purpose GPUs or CPUs ([11] www.researchgate.net). Google pioneered this trend by designing the **Tensor Processing Unit (TPU)**, a custom ASIC tailored to tensor-multiply and accumulate operations. First deployed internally around 2015 (with the v1 TPU unveiled in 2017 ([1] www.researchgate.net)), TPUs initially accelerated inference for Google services. They later gained training capability (TPUv2 in 2017), and have since advanced through multiple generations with steadily increasing performance, memory, and interconnect bandwidth.

This hardware evolution parallels Google's software advances in machine learning. Google introduced the TensorFlow framework (2015) and its specialized XLA compiler early on, laying the groundwork for efficiently mapping neural network graphs to TPU hardware. In 2023, Google introduced the **AI Hypercomputer** architecture, a fully-integrated system encompassing TPU hardware, custom network fabrics (Jupiter, including optical switches), and optimized software (XLA, JAX, PyTorch/XLA). This "vertically integrated" stack is exemplified by Google's own flagship language model series, *Gemini*. Google's Gemini models (built on its DeepMind research) represent its flagship LLM initiative, competing with OpenAI's GPT models. The latest Gemini 3 model (released Nov 2025) boasts a 1 million-token context window, highly improved reasoning, and multimodal capabilities ([12] www.tomsguide.com). Crucially, Google has trained and deployed Gemini 3 entirely on its own TPUs. As one industry analysis observes, "Gemini 3 was developed entirely on [Google's] proprietary TPUs" rather than GPUs, marking a strategic shift away from Nvidia's hardware dominance ([13] www.linkedin.com).

In this report, we delve deeply into the technical underpinnings of Google's TPU accelerators and how they enable Gemini 3's capabilities. We begin by tracing the TPU hardware evolution (generations v1–v7) and key architectural innovations. We then explore the TPU "stack" including chip design, memory systems, interconnection networks, and software layers. We present performance and efficiency metrics (benchmarks, FLOPS/Watt, energy usage) from published studies. Case studies—from Google's own results to customer testimonials—show how TPUs deliver real-world impact. Finally, we discuss the broader implications: how Google's TPU strategy challenges GPU incumbents, influences AI compute economics, and sets a course for future AI accelerators. All sections are supported by citations to peer-reviewed papers, Google technical blogs, news analyses, and conference reports.

TPU Generations: Evolution and Specifications

Google's TPUs have evolved from an inference-focused ASIC to a family of supercomputer-scale accelerators. Table 1 summarizes key characteristics of each generation.

TPU Gen	Announced	Application	Peak Perf. (per chip)	Memory (HBM)	Interconnect & Pod	Notable Features
TPU v1	2015–2017	Inference only	92 TOPS (8-bit) ([1] www.researchgate.net)	28 MB on-chip ([14] www.researchgate.net)	PCIe board, host-managed, up to 16–32 chips per machine	256×256 CPU-style systolic array; fixed schedule; very high TOPS/W ([15] www.researchgate.net). </current_article_content>Deterministic model favored worst-case latency ([16] www.researchgate.net).
TPU v2	2017	Training + Inference	~45 TFLOPS (bfloat16) per chip ([17] gigazine.net)	~8 GB HBM per chip	4 chips per board (~180 TFLOPS each) ([17] gigazine.net); pods up to 64 boards (=11.5 PFLOPS) ([18] gigazine.net)	Introduced bfloat16 training, liquid cooling. Each board (4 chips) 180 TFLOPS, pod 11.5 PFLOPS ([17] gigazine.net) ([18] gigazine.net). Transitioned Google to hardware training.
TPU v3	2018	Training	420 TFLOPS (bfloat16) per 4-chip board ([2] www.slideshare.net)	128 GB per board ([2] www.slideshare.net)	Pods of 1024+ chips (e.g. 2D torus network); >100 PFLOPS at 1024 chips ([3] www.slideshare.net)	Further liquid cooling, higher memory. TPU v3 pods achieved record MLPerf training times (e.g. 16–28 s on common models) ([19] arxiv.org). Depth-first MPI-like scaling.
TPU v4	2020	Training + Inference	~1 petaflops (bfloat16) per chip (est.) ([5] arxiv.org)	≥128 GB per board	**Optical Circuit Switches (OCS)** for pod interconnect ([4] arxiv.org); up to 4096-chip supercomputer.	Architecture built around 5-D torus + re-configurable optical fabric ([4] arxiv.org). Adds **SparseCores**: ~5% of die specialized for embedding layers (5×–7× speedup on embedding-heavy models ([20] arxiv.org)). ~2.1× higher perf than TPU v3 ([21] arxiv.org), and 2.7× better perf/W.
TPU v5e/p	2023	Inference (v5e); Training (v5p)	393 TOPS (int8) per chip ([22] cloud.google.com)	N/A (chip-local, depends on uses)	256-chip pod = 100 PηOPS int8 ([22] cloud.google.com) (100 Peta-OPS). XLA/JAX stack.	Two variants: v5e (inference-optimized, eight "shapes" supporting up to 2T parameter models ([23] cloud.google.com)) and v5p (training-oriented). v5e delivers up to **2.5×** inference throughput/$ over v4, and **1.7×** lower latency ([24] cloud.google.com) ([25] cloud.google.com). Supports int8 quantization for LLMs ([22] cloud.google.com).
TPU v6e (Trillium)	2024 (GA)	Training + Inference	*×4.7*↑ peak per chip vs v5 ([26] cloud.google.com)	*2× v5 HBM capacity* ([26] cloud.google.com)	Jupiter fabric: 100K chips/pod @13 Pbps ([27] cloud.google.com). GKE/AI Hyperform.	Trained Google Gemini 2.0 ([28] cloud.google.com). Major leap: **4.7×** peak perf per chip over v5, **2×** memory and interface bandwidth, **67%** energy efficiency gain ([26] cloud.google.com). Enables 100K-chip pods. Up to **2.5×** better $/training-LLM than v5 ([29] docs.cloud.google.com). Host memory offload to supplement HBM ([27] cloud.google.com).
TPU v7x (Ironwood)	2025 (Preview)	Inference (LLM tuning)	~4,614 TFLOPS per chip (rumored) ([30] medium.com)	192 GB per chip (rumored)	~9,216 chips/pod (inference scale-up); bidirectional interconnect ~1.2 Tbps per pod ([30] medium.com)	First TPU "inference-first" design ([31] medium.com). Analyst projections (unconfirmed) suggest **~4.6 PFLOPS** per chip and **42.5 EFLOPS** per pod ([30] medium.com), with exceptional efficiency (~30× original TPU power efficiency ([32] medium.com)). Still under wraps, but aimed at supporting huge context multimodal models.

Table 1: Summary of Google TPU generations. ("Perf" is peak throughput; TOPS = 10^12 ops/sec. TPU v1 was 8-bit inference-only; later versions use bfloat16 for training BFLOPS or int8 for inference as noted. Citations give sources for key numbers.)

Early on, TPUs already dramatically outperformed general-purpose chips. The original TPU (v1) was a 65,536–multiplier array (256×256) clocked to ~92 TOPS in 8-bit mode ([1] www.researchgate.net). Because it eschewed complex features (no out-of-order exec, only simple buffers), it achieved **15–30×** higher throughput than contemporary GPUs/CPUs on neural network inference, and **30–80×** higher TOPS/Watt ([15] www.researchgate.net). Thus even the first TPU gave orders-of-magnitude energy savings on Google's datacenter inference tasks ([15] www.researchgate.net).

TPUv2 in mid-2017 was the first to add training capability. It packed *four* TPU chips per board and 8 GB HBM each, delivering ~180 TFLOPS per board in bfloat16 precision ([17] gigazine.net). A full TPU v2 "Pod" of 64 boards achieved roughly 11.5 petaflops of peak training throughput ([18] gigazine.net). Subsequent TPUv3 scaled this further: each v3 board (eight high-end chips) provided ~420 TFLOPS of bfloat16 compute and 128 GB of HBM ([2] www.slideshare.net), and a 1024-chip pod could exceed 100 petaflops with ~32 TB of aggregate memory ([3] www.slideshare.net). Notably, Google's MLPerf results used TPU v3 pods to set world-record training times (e.g. training ResNet-50 in 16 seconds) ([19] arxiv.org).

TPU v4 (2020–21) introduced revolutionary interconnect. It embeds **optical circuit switches (OCS)** in the pod network, allowing dynamic reconfiguration of a 5-dimensional torus fabric ([4] arxiv.org). These OCS links have very low latency and power: they cost <5% of a pod's budget and consume <3% of its power ([4] arxiv.org), yet enable scaling out to ~4096 chips in Google Cloud setting. Figure-wise, TPU v4 offers roughly **2.1×** the raw performance of TPU v3 on typical ML workloads, and **2.7×** better performance-per-watt ([21] arxiv.org). The v4 die also includes specialized *SparseCores* for embedding tables: these small dataflow units (only ~5% of the silicon) accelerate sparse lookup/embedding-intensive models in NLP and recommendation by ~5–7× ([21] arxiv.org).

Cloud TPU v5 (2023) split into two lines: **v5p** for training and **v5e** for inference. The v5e chip boasts an enormous 393TOPS of INT8 compute (about 0.393 peta-ops) ([22] cloud.google.com). Its 256-chip pod reaches ~100 peta-ops (INT8), allowing large LLMs (up to ~2 trillion parameters) to be served efficiently ([22] cloud.google.com) ([33] cloud.google.com). Google reports that optimized Stack (XLA, quantization) yields up to **2.5×** better inference throughput-per-dollar and **1.7×** lower latency on v5e vs. v4 ([24] cloud.google.com) ([25] cloud.google.com). For training, the v5p variant provided roughly double the HBM capacity of v4 boards (exact figures undisclosed) and other incremental gains.

The **6th-generation (TPU v6e "Trillium")** was GA in late 2024 ([34] docs.cloud.google.com). Trillium more than quadruples per-chip compute (x4.7 peak vs. v5) while doubling memory and interconnect bandwidth ([26] cloud.google.com). It is described as the foundation of Google's *AI Hypercomputer*: >100,000 chips can be linked with >13 petabit/s bisectional bandwidth on the Jupiter network ([27] cloud.google.com). Trillium's efficiency is also dramatic: roughly a 67% increase in TOPS/Watt over v5 ([26] cloud.google.com). Google explicitly credits Trillium TPUs for training Gemini 2.0 (the predecessor to Gemini 3) ([28] cloud.google.com). In large LLM training, Trillium can be 2.1–2.5× more cost-effective than v5 pods ([29] docs.cloud.google.com).

Finally, Google has previewed **TPU v7x ("Ironwood", 2025)** for ultra-large-scale AI. Details remain proprietary, but industry estimates are staggering: one analysis projects ~4,614 TFLOPS (4.6 petaFLOPS) per chip of mixed-precision compute and ~192 GB HBM ([30] medium.com), scaling to ~42.5 exaFLOPS in a 9,216-chip pod ([30] medium.com). Ironwood is said to be "inference-first" rather than general-purpose ([31] medium.com), targeting trillions of real-time queries (RDNA classes of workload). If these projections hold, TPU v7 would deliver orders of magnitude more throughput than a typical GPU cluster, while maintaining **~30×** the efficiency of the original TPU generation ([32] medium.com).

Each generation of TPU thus roughly maintains the trend of year-over-year leaps in throughput and Watt-performance that Moore's Law in general-purpose CPUs no longer guarantees. The evolution from an initial 92 TOPS inference chip ([1] www.researchgate.net) to projected **4,600 TFLOPS** training/inference chips ([30] medium.com) over ~8 years reflects focused ASIC specialization. Table 1 encapsulates these changes.

Technical Architecture of Google TPUs

At the heart of every TPU is a large *systolic array* optimized for matrix multiplication and convolution, the core operations in neural networks. For example, the original TPU v1 contained a 256×256 array of 8-bit multiply-accumulate units, yielding the 65,536 MACs that delivered its 92 TOPS peak ([1] www.researchgate.net). Later TPUs continued this pattern: v2 boards had 4 chips (each ~45 TFLOPS BF16), and v3 boards 8 chips (~420 TFLOPS BF16) ([17] gigazine.net) ([2] www.slideshare.net). By TPU v7, the array size is not public, but the conjectured 4.6 PFLOPS peak suggests on-chip MAC counts in the billions. Crucially, TPUs omit general-purpose logic: they have no out-of-order execution or deep cache structures. Instead, operations are scheduled in a static, deterministic pipeline, trading flexibility for efficiency. This design choice yields very predictable 99th-percentile latency and exceptionally high utilization of the MAC units ([16] www.researchgate.net).

Alongside the array, each TPU chip includes a sizeable on-chip memory (Unified Buffer) and a "weight FIFO" feeding weights from off-chip memory into the array ([35] www.researchgate.net). For inference, recent TPUs incorporate INT8 (and INT4/BF4) arithmetic to

maximize throughput with minimal loss in model accuracy; for example, TPU v5e achieves 393 TOPS performing INT8 ops ([22] cloud.google.com). For training and sensitive computations, TPUs support the 16-bit Brain Float (bfloat16) format, which provides a wide dynamic range. Early TPUs used 8-bit for inference, but v2 and later defaults to bfloat16 for all MAC operations.

A key strength of TPUs is their high-bandwidth memory system. TPU v3 boards include up to 128 GB of High-Bandwidth Memory (HBM) per 4-chip board ([2] www.slideshare.net). The HBM stacks are directly attached via wide buses to feed the MAC arrays at full speed, enabling large models to run across many chips without off-chip memory stalls. In TPU v4/v5, Google doubled and again doubled the memory bandwidth per chip compared to prior gens ([36] docs.cloud.google.com) ([26] cloud.google.com). Externally, chips are linked by the Jupiter network fabric: a high-radix, packet router–based network that, in current deployments, connects up to 100,000 TPU chips with a bisectional bandwidth of ~13 petabit/s ([27] cloud.google.com). The fabric supports collective operations (broadcast, reduce) essential for distributed training.

Significantly, TPU v4 introduced *optical* switching into this network. Unlike typical electrical switch fabrics (e.g. InfiniBand), Google's TPU pods employ optical circuit switches (OCS) that can dynamically reconfigure connections between racks ([4] arxiv.org). In practice, this means that subsets of TPUs can be wired together in flexible topologies (3D torus, etc.) optimized for a given job. The OCS equipment itself is low-cost and low-power (<5% of the pod's cost/power ([4] arxiv.org)), dramatically reducing network overhead while improving effective bandwidth. This innovation was key to enabling TPU v4 pods of **4096 chips** (~10× larger than v3 pods) ([37] arxiv.org).

Another TPU-specific unit is the **Sparse Core** (on v4 and later). Sparse Cores are small specialized processors designed for embedding table lookups and other sparse/dense computations that do not map efficiently to the big dense array. Each TPU v4 includes such a Sparse Core that can accelerate embedding-laden models by ~5–7× while consuming only ~5% of the chip area and power ([21] arxiv.org). This is particularly useful for recommendation systems or large LMs that use categorical embeddings (e.g. large vocabularies, one-hot).

On the software side, Google tightly integrates TPUs with its ML stack. The XLA compiler targets TPU instruction sets, partitioning tensor graphs across multiple chips/workers. Google's Pathways and JAX frameworks natively pace work onto TPU pods, allowing single models to spread over thousands of chips when needed. This synergy is evidenced by Gemini 3: internal docs note that its training ran on TPU v5e/v6e pods combined with XLA/JAX/Pathways to minimize compute fragmentation ([6] www.linkedin.com). TPU Runtime handles I/O, host offloading (Trillium allows massive host RAM to supplement HBM ([27] cloud.google.com)), and collective communication. The entire TPU ecosystem is built to ensure that "what you train is what you serve": the same chip and precision used in training can be used for low-latency inference (e.g. TensorFlow, PyTorch/XLA, Vertex AI direct-support).

In summary, each TPU generation co-designs chips, board interconnect, and compiler to maximize throughput for neural nets. The large matrix arrays supply the bulk of compute; wide HBM and fabric ensure data keeps up; specialized units (like Sparse Cores) and optimized numeric formats (bfloat16, INT8, FP8 in future) accelerate common layers; and software (XLA/JAX) squeezes out efficiency. This full-stack optimization is why TPUs can beat GPUs on large AI workloads despite comparable or even lower raw transistor counts.

Performance, Scalability, and Efficiency

Google's TPUs consistently demonstrate leading performance and efficiency on AI benchmarks. In the seminal TPU v1 evaluation, Google reported that even at low utilization, the TPU was **15×–30×** faster on neural-net inference workloads than contemporary Intel CPUs or Nvidia K80 GPUs, with **30×–80×** higher TOPS/Watt ([15] www.researchgate.net). In practical terms, replacing old GPUs/CPUs in Google's datacenter with TPUs reduced inference latency by an order of magnitude and power draw by similar factors.

Comparing more recent chips, Jouppi *et al.* (2023) find that TPU v4 pods far outperform similar-size alternatives. For example, a TPU v4 supercomputer (4096 chips) is **4.3–4.5×** faster than a Graphcore IPU Bow (a dedicated AI chip system) of comparable size, and **1.2–1.7×** faster than an Nvidia A100 pod ([5] arxiv.org). Importantly, the TPU v4 achieves this while using **1.3–1.9×** *less* power than the A100 pod ([5] arxiv.org). In on-premise comparisons, shifting a workload onto Google Cloud TPU v4 reduced energy usage by ~3× and CO_2 emissions by ~20× relative to local GPU servers ([5] arxiv.org). Table 2 presents a few illustrative comparisons.

System	Relative Throughput	Relative Power	Notes
TPU v1 (2017) vs. contemporaneous GPU	15–30× higher on NN inference ([15] www.researchgate.net)	30–80× higher TOPS/Watt ([15] www.researchgate.net)	Measured on typical Google inference tasks. TPU achieved far better speed and efficiency.
TPU v4 (2020) vs. Nvidia A100 outlet	~1.2–1.7× higher overall ([5] arxiv.org)	Uses ~53–77% of A100 power ([5] arxiv.org)	For similarly-sized pods (chips count) on real LLM training tasks.
TPU v4 vs. Graphcore IPU	~4.3–4.5× higher ([5] arxiv.org)	—	Peer chips scaled to comparable cluster size.
TPU v5e vs. TPU v4 (inference)	Up to 2.5× perf/$, 1.7× lower latency ([24] cloud.google.com) ([25] cloud.google.com)	—	Google reports TPU v5e gives 2.5× throughput per dollar vs v4 on LLM inference.

System	Relative Throughput	Relative Power	Notes
TPU v6e (Trillium) vs. TPU v5	4× training perf↑, 3× inference↑ ([36] docs.cloud.google.com)	+67% efficiency ([26] cloud.google.com)	Cloud TPU release notes list Trillium improvements.

Table 2: Performance and efficiency comparisons of Google TPUs with alternatives or prior generations. "Relative Throughput/Power" gives approximate factors from cited sources.

These published figures confirm that TPUs deliver competitive (often superior) performance on AI benchmarks. For machine learning contests like MLPerf, Google has used TPU pods to win top scores. For instance, a Google report shows a TPU v3 multipod (4096 chips) trained MLPerf models (ResNet-50, GNMT, SSD, etc.) in record times (16–28 s) by employing advanced model/data parallelism ([19] arxiv.org). Later, Google announced that TPU v4 pods drove MLPerf training to even faster ceilings, and that TPU-based models often sweep ML benchmarks for large-scale LMs and reasoning tasks ([38] www.linkedin.com) ([39] www.itpro.com).

Piercing the hype, independent analyses also highlight TPU eco-nomics. Google's own TPU pricing model is predictable: e.g. $2.50 per million input tokens for Gemini 2.5 (and similar rates for Gemini 3), reflecting the fixed GPU-nightly cost structure ([40] expertbeacon.com). But more deeply, industry letters (Stanford's Crux project, etc.) note that horizontally integrated TPU infrastructure yields lower amortized $/TFLOP than GPU systems. In particular, one detailed study found that a fully-utilized TPU Pod can train an LLM at noticeably lower capex and opex than an equivalent Nvidia H100 cluster ([8] www.linkedin.com). This stems from the higher FLOPS/W and stability of TPU-based systems: Google states that v5e/v6e chips produce *predictable* cost-per-token, avoiding the load spikes common in bespoke GPU fleets ([41] www.linkedin.com).

We also consider power and sustainability: as demand for exascale training grows, energy per model has become critical. Google's reports emphasize that TPU-based data centers (especially TPU v4+ in custom "Hypercomputer" design) are far more energy-efficient for AI. **Google claims** that a Google Cloud TPU v4 deployment uses ~3× less electricity and emits ~20× less CO_2 than a standard on-prem cluster doing the same training ([42] arxiv.org). Meanwhile, independent commentators (e.g. Prakash 2025 ([32] medium.com)) project that next-gen TPUs (like Ironwood) could be ~30× more efficient than v1, vastly outstripping GPUs' gains in performance/Watt. These levels of efficiency grant organizations running massive inference workloads (billions of queries/day) a tangible cost sinks in the hundreds of millions of dollars.

In practical terms, many AI companies have switched to TPUs for their high-volume services. AssemblyAI, which does massive speech-recognition inference, reports that Cloud TPU v5e gave "up to 4× greater performance per dollar" versus other solutions, dramatically speeding their deployment ([9] cloud.google.com). Similarly, the conversational AI company Gridspace cites **5× training speedups** and **6× larger inference scale** on Google TPUs for their models ([10] cloud.google.com). AI21 Labs (developer of LLMs) has long benchmarked on TPUs and now uses the Trillium TPU for its Mamba/Jamba language models, explicitly noting that the "advancements in scale, speed, and cost-efficiency are significant" ([43] cloud.google.com). These vignettes show that real users validate TPMs: on complex AI tasks, TPUs can simplify workflows (optimized kernels) reduce engineering effort, and above all deliver more compute for the money.

TPUs versus GPUs and Other Accelerators

In the broader AI hardware landscape, Google's TPU strategy stands in contrast to the GPU-centric paradigm. For the last decade, Nvidia dominated the datacenter AI accelerator market (85–90% share ([44] www.linkedin.com)) with versatile GPU families (A100, H100, and next-generation Blackwell/B200). GPUs excel in generality (supporting CUDA, OpenCL, PyTorch, TensorFlow, etc.), but their design isn't specialized for deep learning. An analyst aptly summarizes: *"GPUs are the Swiss Army knife of computing; TPUs are a scalpel for AI."* ([45] medium.com). This distinction means TPUs can achieve higher AI-specific efficiency at scale, at the expense of broader applicability.

Nvidia's latest Hopper (H100) and upcoming Blackwell (B200) GPUs are formidable: H100 has ~3,958 TFLOPS (with FP8) and Blackwell promises new low-precision modes (FP4/FP6) and 192 GB HBM3e ([46] medium.com). These GPUs offer robust ecosystems (CUDA, cuDNN, TensorRT) and support for everything from scientific computing to gaming. However, their maximum AI performance-per-watt is bounded by their general-purpose architecture. In Dilworth's words, *"this versatility comes at a power consumption cost."* ([47] medium.com).

By contrast, Google's TPU roadmap is fixed on AI. Each TPU gen trades broad flexibility for focused acceleration. For instance, Nvidia's FP8 support yields huge TFLOPS, but only on certain arithmetic, while TPU's entire pipeline (including sparse units and compiler) is optimized for neural graph execution. In practice, as Google showed, training LLMs entirely on TPUs (v3–v6) can match or exceed what one would get from a walled garden of tens-of-thousands of GPUs ([13] www.linkedin.com).

Google's shift to TPU-only ("TPU revolution") for Gemini 3 is historically significant. It is reported that **all phases** of Gemini 3 training ran on Google-made TPU v5e and v6e pods without fallback to Nvidia GPUs ([6] www.linkedin.com). Early Google frontier models often used a mix of GPUs and TPUs, but Google now claims its large models need only its ASIC pipeline. This move upends expectations: previously, AI startups and labs largely lined up to lease Nvidia GPU clusters (e.g. OpenAI, Meta, Anthropic). Now Google demonstrates that self-built accelerators can credibly compete. In one analysis, Gem3 was described as "the first real challenge to Nvidia's dominance" because it broke the de facto Nvidia monopoly on *training* chips ([13] www.linkedin.com).

Nevertheless, it is worth noting some caveats. TPUs remain somewhat proprietary within Google's ecosystem. They require using Google Cloud or Google hardware environments (Vertex AI, GKE) and supporting frameworks (TensorFlow, JAX, PyTorch/XLA). Some customers worry about vendor lock-in or the difficulty of porting existing workloads to TPU's more rigid model. In practice, broad frameworks like PyTorch now have TPU backend (PyTorch/XLA), but the ecosystem is less mature than CUDA's. On multi-cloud or on-prem, fewer organizations as-of-2025 run pure TPUs. For mixed or legacy workloads, GPUs (especially with cuDNN/TensorRT) still have an edge in flexibility and library support.

From the perspective of large-scale deployers (hyperscalers, Google itself, AWS, etc.), however, the price-performance trade-offs are shifting. Some analysts note that for massive inference workloads, Google's TPUs (v7 Ironwood) and AWS's own silicon (Trainium/Inferentia) are poised to dominate price-performance, while GPUs might be favored only for the most general tasks ([48] medium.com) ([49] medium.com). In short, Google's TPUs appear to capture an advantage for "production AI at scale" (banking, search, social, mapping), whereas GPUs remain popular for research flexibility and niche compute.

In the academic/public sphere, comparative studies have reinforced Google's claims. Independent benchmarks (e.g. ServeTheHome) confirm the published Google data: large TPU pods consistently meet or beat GPU clusters on raw training tasks, while GPUs shine on other mixed workloads or single-card tasks. For example, AI21 Labs (customer testimonial) explicitly moved their LLM training to Google TPUs after earlier multi-year use of Nvidia GPUs, citing substantial cost/perf wins ([43] cloud.google.com). The result is an increasingly multipolar chip market: Nvidia is no longer the sole leader — Intel (Gaudi CPUs), AMD (Instinct MI300), Graphcore IPU, and even specialized PIM chips (e.g. SambaNova, Cerebras) all contend. But Google's end-to-end control — hardware, software, data centers — means its TPU offerings are uniquely integrated, and *that* vertical stack is what powers Gemini 3's breakthroughs.

TPUs Powering Gemini 3: Case Study

The Gemini 3 large language model exemplifies Google's TPU-centric strategy. Released November 18, 2025, Gemini 3.0 is a multimodal LLM with an unprecedented 1-million-token context window and advanced reasoning ("Deep Think" mode) ([12] www.tomsguide.com). Such a model demands immense computation and memory. According to publicly available information and expert analysis, Google trained and serves Gemini 3 exclusively on its latest TPUs (v5e and v6e) under the hood ([6] www.linkedin.com) ([7] www.itpro.com). This has several technical implications:

- **Scale of computation:** Gemini 3 reportedly employs a "Mixture of Trillions" architecture—perhaps 2–4 trillion total parameters with sparse activation of ~150–200B per query ([50] expertbeacon.com). Training such a massive model requires distributed computation across many accelerators. Google's TPUv4 Pod (4096 chips) and now TPUv5e/v6e pods (with optical DCN) provide the needed scale and HW support. The LinkedIn analysis notes that Google's experience training Gemini Ultra (a component of Gemini 3) on multisite TPUv4 clusters laid the operational foundation for scaling to trillions of parameters ([51] expertbeacon.com). In other words, Gemini 3 leverages TPU pods that span multiple data centers, synchronized with Pathways.
- **Large memory and embedding support:** A 1M-token context means the model must ingest and attend over huge sequences. This in turn requires vast memory bandwidth. TPUs supply this: for example, each TPU chip now has significantly more HBM (especially v6e, double v5 ([26] cloud.google.com)). Moreover, Gemini 3's multimodality (text+image+code+audio) and embedding layers are accelerated by TPU's **SparseCore** units and unified memory, which excel at large sparse lookups. Qualcomm's analysis highlights that TPU v4's embedding cores can give 5–7× speedups for such layers ([21] arxiv.org), crucial when multimodal data often involve large dictionaries or vision feature maps.
- **Real-time video and 3D:** Technical speculation (e.g. ExpertBeacon ([52] expertbeacon.com)) suggests Gemini 3 might support real-time (60 FPS) video understanding and native 3D spatial reasoning. If true, this is only possible with hardware that can process tens of billions of operations per second in a streaming fashion. TPU architecture is well-suited: the dataflow array and massive inter-chip fabric could continuously stream frames into the model. Expert analyses note that achieving 60 FPS inference would require *explicit* temporal attention mechanisms and highly optimized kernels (perhaps custom XLA kernels running on TPU) ([53] expertbeacon.com). While we haven't seen official confirmation, it underscores that Google believes TPUs can handle next-gen AI workloads beyond text.
- **Embedded reasoning (Deep Think):** Gemini 3 removes the user-visible "Deep Think" toggle by presumably integrating iterative reasoning loops directly in the model. That likely means Gemini 3 runs *internal verification and refinement passes* when processing complex queries. This aligns with TPU strengths: the hardware supports streaming repeated inference steps efficiently. In Gemini 2.5, Deep Think was a slower mode; now, Table [expert] suggests (Alex Fello analysis) that Gemini 3 embeds a fast verifier network to trigger extra compute when needed ([54] expertbeacon.com). Each such loop in the network becomes a TPU sub-inference. The compute cost can scale out over many TPU slices, with the low-overhead NIC fabric and parallel TPU pipeline hiding some latency. Google's design likely uses TPU's high memory capacity to maintain context between these loops.
- **Training regime:** The LinkedIn source confirms Google trained Gemini 3 on TPU v5e and v6e pods ([6] www.linkedin.com). This means every weight update (backprop through billions of parameters) happened on Google chips. By contrast, models like GPT-4 used ExaTensor or Nvidia clusters. Google's switch to TPU-only means it did not rely on external hardware constraints. It also implies Google needed advanced optimizations: things like 2D/3D mesh partitioning, ORCA work stealing, or the Bagua/JAX-allreduce stack to keep the high-throughput TPU pods fully utilized. We can infer that training Gemini 3 took orders of magnitude more aggregate compute than prior LMs, but TPU specialization likely compressed the timeline by yielding more TFLOPS per chip.

- **Latency and inference:** Gemini 3's deployment (e.g. integrated into Search and Google Workspace) benefits from TPU inference. The new "Flash" variant of Gemini 3 (released Dec 2025) emphasizes speed and low cost, claiming latency akin to simple search queries ([55] www.axios.com). Such claims rely on TPU inference pods. The TPU fabric's bi-directional bandwidth and the sophisticated software stacks (serving on VertexAI or Antigravity IDE) ensure that a user's query is offloaded to an accelerator very quickly, processed with the full model, and returned in sub-second time. Google's assertion that Gemini 3 Flash is both faster and cheaper than Gemini 3 Pro ([56] www.axios.com) strongly implies that v7/Trillium-level hardware is delivering realtime performance with lower cost per query.
- **Benchmark performance:** External benchmarks corroborate TPU effectiveness. For example, Google published that Gemini 3 Pro scored 1501 Elo on LanguageModelArena (a text-agility test) and 91.9% on GPQA-Diamond (PhD-level QA), among the best ever ([38] www.linkedin.com). These results came from the TPU-trained model, validating that the hardware allowed achieving frontier capabilities. Comparisons show Gemini 3 Pro outperformed contemporaries like GPT-5.1 and Anthropic's models across many tasks ([57] www.itpro.com). Gemini 3's specialized Deep Think mode even hit 41% on the "Humanity's Last Exam" test (very hard reasoning) ([58] www.itpro.com). The key enabler of these capabilities was TPU: the ItPro report explicitly credits Google's own TPUs for Gemini 3's training and inference efficiency ([58] www.itpro.com).

In summary, Gemini 3 serves as a **case study** illustrating how Google's TPU supercomputers realize cutting-edge AI. Every aspect of the model—from its trillion-scale sparsity to interleaved reasoning loops—aligns with TPU strengths (massive parallelism, high memory, and custom logic). Google's internal and external accounts unanimously affirm: *"Gemini 3's training was powered by our AI-specific TPUs"* ([58] www.itpro.com) ([6] www.linkedin.com). This tight hardware–software co-design has tangible effects: it enabled Gemini 3 to leapfrog competitors on benchmarks, deliver new multimodal features, and scale to unprecedented context lengths, all while containing cost and energy.

Strategic and Future Implications

Google's TPU-centric approach has broad implications for the AI ecosystem. On the business side, it challenges Nvidia's hegemony in AI compute. With TPU-exclusive training of Gemini 3, Google proves that a top-tier LLM can be developed without GPUs. Industry analysis has taken note: by eliminating reliance on Nvidia GPUs, Google "alters the cost-to-performance economics" of training frontier models ([13] www.linkedin.com). The LinkedIn report estimates that fully-utilized TPU v5/v6 pods have lower capital expenditure per FLOP than equivalent H100 clusters ([8] www.linkedin.com). If Google can maintain this lead, it may force others to either invest in their own ASICs (e.g. AWS Trainium, Meta's plans) or pay a premium for limited GPU supply.

From a compute economics perspective, Google claims an ongoing advantage. Each new TPU gen has delivered a large jump in price-performance (e.g. 2.1–2.5× better $/performance for training LLMs ([29] docs.cloud.google.com)). Moreover, because TPUs are tightly integrated with cloud infrastructure, Google can amortize costs over massive scale (Google claims tens of exaflops on the Jupiter network ([27] cloud.google.com)). One analysis notes that these efficiencies "compound into tangible operational advantages" for hyperscalers handling billions of inference requests daily ([48] medium.com). In other words, if Google's TPU-driven Gemini 3 gains translate into lower costs per query, it could squeeze margin from competitors.

The environmental impact is also key. By using much less energy for the same AI workload, TPUs support more sustainable AI. Google reports that its TPU-based data centers (especially "AI Hypercomputer" designs with v4+) emit far less CO_2 per task ([5] arxiv.org). This may become a competitive necessity as corporate and regulatory pressures mount to reduce the carbon footprint of AI. Per [5], TPU v4 uses ~3× less energy and generates ~20× less carbon-equivalent than a comparable GPU cluster ([5] arxiv.org), an enormous gap. As models like Gemini 3 proliferate into consumer products, energy cost becomes a strategic factor; TPUs help Google claim a sustainability advantage in AI.

Looking ahead, the evolution of TPUs (into v7 and beyond) suggests Google is committing to this path. The upcoming TPU7x "Ironwood" (7th gen) underscores a continued race: it is already described as an "inference-first" chip with unheard-of throughput ([31] medium.com). If Google's descriptions bear out, Ironwood will allow **massive parallel inference** on Gemini-like models across many chips, with unprecedented energy efficiency. Concurrently, Google is enhancing software (e.g. expanding PyTorch/XLA support, improved XLA compiler optimizations ([59] cloud.google.com)) to make TPUs more accessible. The interplay of these developments may lock in Google (and its cloud customers) even more deeply into the TPU ecosystem.

Of course, multiple perspectives exist. Some analysts caution that relying on TPU locks one into Google's stack; broad workloads might still prefer GPUs or specialized accelerators from AMD/Meta. Notably, AMD has committed to open standards (ROCm) and is investing in AI chiplets (e.g. MI300 series). Meanwhile, AWS continues to differentiate with its own silicon (Tranium/Inferentia) focusing on cost-effective inference ([48] medium.com). From the vantage of enterprises, it remains to be seen how much benefit out-of-Google companies get from TPUs versus proven GPU ecosystems. NVIDIA is responding with its own innovations (e.g. H200 "Blackwell" with low-bit precision and chiplet designs ([46] medium.com)), so GPUs will remain competitive, especially where NVIDIA's software stack and familiar APIs are preferred.

In academia and industry, there is excitement about TPU's architectural ideas spreading. For example, optical switching and sparsity accelerators pioneered in TPU v4 are being explored in wider data-center designs (even outside of Google). The notion of "AI Hypercomputer" as a unified hardware/software paradigm may influence other players. Case in point: the Open Compute Project launched an Optical Circuit Switching initiative in 2023, acknowledging Google's success with OCS. Likewise, the TPU's emphasis on high-TOPS/W

accelerators may spur more ASIC development for AI: Intel's Gaudi/Habana chips, Graphcore's IPU, and even startups with PIM architectures are partly reactions to TPU-style specialization.

In summary, Google's use of TPUs — especially to underpin Gemini 3 — signals a shift. The foreseeable future of large-scale AI compute is increasingly heterogeneous: a mix of cloud ASICs (like TPUs), GPUs, and domain-specific chips. Google (and collaborators like Meta, Amazon, etc.) are broadly heading that direction. The advantage is clear for now: Google's in-house TPUs deliver top-tier model performance, excellent energy efficiency, and beneficial economics ([8] www.linkedin.com) ([5] arxiv.org). The challenge will be to maintain flexibility as AI systems evolve. Can the benefits of specialized hardware keep pace with the rapidly-changing demands of AI? For Gemini 3, at least, the answer so far appears to be **yes**: Google's TPU "powerhouse" has enabled a generation of model capabilities that might have been far costlier or slower to achieve on any other hardware.

References

- Jouppi, N. P., Young, C., Patil, N., & Patterson, D. (2017). *In-datacenter performance analysis of a tensor processing unit*. *Proceedings of the 44th Annual International Symposium on Computer Architecture*, pp. 1–12. DOI:10.1145/3079856.3080246 ([1] www.researchgate.net) ([15] www.researchgate.net).
- Jouppi, N. P., Kurian, G., Li, S., Ma, P., Nagarajan, R., Nai, L., … Patterson, D. (2023). *TPU v4: An Optically Reconfigurable Supercomputer for Machine Learning with Hardware Support for Embeddings*. (ArXiv preprint) ([4] arxiv.org) ([5] arxiv.org).
- Kumar, S., Kumar, N., Lee, H., et al. (2019). *Scale MLPerf-0.6 models on Google TPU-v3 Pods*. (ArXiv:1909.09756) ([60] arxiv.org).
- Kumar, S., Bradbury, J., Young, C., et al. (2020). *Exploring the limits of concurrency in ML training on Google TPUs*. (ArXiv:2011.03641) ([19] arxiv.org).
- Spiridonov, A. & Ji, G. (2023). *How Cloud TPU v5e accelerates large-scale AI inference*. Google Cloud Blog, Aug 31, 2023 ([24] cloud.google.com) ([9] cloud.google.com).
- Lohmeyer, M. (2024). *Announcing the general availability of Trillium, our sixth-generation TPU*. Google Cloud Blog, Dec 11, 2024 ([59] cloud.google.com) ([26] cloud.google.com).
- Google Cloud Platform. *Cloud TPU release notes* (including Trillium announcement). Google Cloud Docs ([34] docs.cloud.google.com) ([61] docs.cloud.google.com).
- Google support. *Cloud TPU Documentation*. Google Cloud (various pages on TPU versions; see release notes).
- Teletchea, S. (2025). "Google's TPU revolution: Why Gemini 3 is the first real challenge to Nvidia's dominance." LinkedIn article, Dec 2025 ([13] www.linkedin.com) ([8] www.linkedin.com).
- *Google's new Gemini 3 Flash is fast, cheap and everywhere.* Axios (Dec 17, 2025) ([56] www.axios.com).
- Huffman, W. (2025). *Google blows away competition with powerful new Gemini 3 model*. ITPro (Nov 18, 2025) ([57] www.itpro.com) ([7] www.itpro.com).
- *Google Gemini 3 – everything you need to know*. Tom's Guide (Nov 20, 2025) ([12] www.tomsguide.com).
- ExpertBeacon AI Analysis (2025). *Gemini 3.0: Technical Analysis of Google's Next-Generation AI Architecture* ([51] expertbeacon.com) ([52] expertbeacon.com).
- "The Great AI Chip Showdown: GPUs vs TPUs in 2025". Medium (Harsh Prakash, Nov 23, 2025) ([31] medium.com) ([30] medium.com) ([32] medium.com).
- *Cloud TPU V4 vs V5e vs V5p Comparison*. ServeTheHome (Dec 2025) – (cited in internal docs).
- AssemblyAI testimonials. (Quoted in [24] Cloud blog) ([9] cloud.google.com).
- Gridspace testimonials. (Quoted in [24] Cloud blog) ([10] cloud.google.com).
- Google DeepMind Blog (Dec 2024). *Trillium TPU is GA*. Google Cloud Blog ([28] cloud.google.com) ([43] cloud.google.com).
- Wikipedia and Google Cloud documentation on TPU (architecture overview).
- Additional industry news and analysis (e.g. TechRadar, Blognone, etc.) for context.

External Sources

[1] https://www.researchgate.net/publication/317613363_In-Datacenter_Performance_Analysis_of_a_Tensor_Processing_Unit#:~:Unit%...

[2] https://www.slideshare.net/grigorysapunov/deep-learning-hardware-landscape#:~:https...

[3] https://www.slideshare.net/grigorysapunov/deep-learning-hardware-landscape#:~:A%20%...

[4] https://arxiv.org/abs/2304.01433#:~:archi...

[5] https://arxiv.org/abs/2304.01433#:~:along...

[6] https://www.linkedin.com/pulse/googles-tpu-revolution-why-gemini-3-first-real-satyajit-chaudhuri-jvnyc#:~:Accor...

[7] https://www.itpro.com/technology/artificial-intelligence/google-launches-flagship-gemini-3-model-and-google-antigravity-a-new-agentic-ai-development-platform#:~:it%20...

[8] https://www.linkedin.com/pulse/googles-tpu-revolution-why-gemini-3-first-real-satyajit-chaudhuri-jvnyc#:~:1,inf...

[9] https://cloud.google.com/blog/products/compute/how-cloud-tpu-v5e-accelerates-large-scale-ai-inference#:~:%E2%8...

[10] https://cloud.google.com/blog/products/compute/how-cloud-tpu-v5e-accelerates-large-scale-ai-inference#:~:%E2%8...

[11] https://www.researchgate.net/publication/317613363_In-Datacenter_Performance_Analysis_of_a_Tensor_Processing_Unit#:~:Many%...

[12] https://www.tomsguide.com/ai/google-gemini-3-everything-you-need-to-know#:~:Googl...

[13] https://www.linkedin.com/pulse/googles-tpu-revolution-why-gemini-3-first-real-satyajit-chaudhuri-jvnyc#:~:in%20...

[14] https://www.researchgate.net/publication/317613363_In-Datacenter_Performance_Analysis_of_a_Tensor_Processing_Unit#:~:throu...

[15] https://www.researchgate.net/publication/317613363_In-Datacenter_Performance_Analysis_of_a_Tensor_Processing_Unit#:~:low%2...

[16] https://www.researchgate.net/publication/317613363_In-Datacenter_Performance_Analysis_of_a_Tensor_Processing_Unit#:~:TPU%2...

[17] https://gigazine.net/gsc_news/en/20170518-google-tpu-2nd-gen/#:~:This%...

[18] https://gigazine.net/gsc_news/en/20170518-google-tpu-2nd-gen/#:~:This%...

[19] https://arxiv.org/abs/2011.03641#:~:a%20s...

[20] https://arxiv.org/abs/2304.01433#:~:Much%...

[21] https://arxiv.org/abs/2304.01433#:~:inclu...

[22] https://cloud.google.com/blog/products/compute/how-cloud-tpu-v5e-accelerates-large-scale-ai-inference#:~:Each%...

[23] https://cloud.google.com/blog/products/compute/how-cloud-tpu-v5e-accelerates-large-scale-ai-inference#:~:,infe...

[24] https://cloud.google.com/blog/products/compute/how-cloud-tpu-v5e-accelerates-large-scale-ai-inference#:~:Up%20...

[25] https://cloud.google.com/blog/products/compute/how-cloud-tpu-v5e-accelerates-large-scale-ai-inference#:~:On%20...

[26] https://cloud.google.com/blog/products/compute/trillium-tpu-is-ga#:~:,trai...

[27] https://cloud.google.com/blog/products/compute/trillium-tpu-is-ga#:~:acros...

[28] https://cloud.google.com/blog/products/compute/trillium-tpu-is-ga#:~:We%20...

[29] https://docs.cloud.google.com/tpu/docs/release-notes#:~:,hund...

[30] https://medium.com/%40hs5492349/the-great-ai-chip-showdown-gpus-vs-tpus-in-2025-and-why-it-actually-matters-to-your-bc6f55479f51#:~:Ironw...

[31] https://medium.com/%40hs5492349/the-great-ai-chip-showdown-gpus-vs-tpus-in-2025-and-why-it-actually-matters-to-your-bc6f55479f51#:~:The%2...

[32] https://medium.com/%40hs5492349/the-great-ai-chip-showdown-gpus-vs-tpus-in-2025-and-why-it-actually-matters-to-your-bc6f55479f51#:~:But%2...

[33] https://cloud.google.com/blog/products/compute/how-cloud-tpu-v5e-accelerates-large-scale-ai-inference#:~:LLMs%...

[34] https://docs.cloud.google.com/tpu/docs/release-notes#:~:This%...

[35] https://www.researchgate.net/publication/317613363_In-Datacenter_Performance_Analysis_of_a_Tensor_Processing_Unit#:~:TPU%2...

[36] https://docs.cloud.google.com/tpu/docs/release-notes#:~:Here%...

[37] https://arxiv.org/abs/2304.01433#:~:TPU%2...

[38] https://www.linkedin.com/pulse/googles-tpu-revolution-why-gemini-3-first-real-satyajit-chaudhuri-jvnyc#:~:Googl...

[39] https://www.itpro.com/technology/artificial-intelligence/google-launches-flagship-gemini-3-model-and-google-antigravity-a-new-agentic-ai-development-platform#:~:to%20...

[40] https://expertbeacon.com/gemini-3-0-technical-analysis-architecture/#:~:Archi...

[41] https://www.linkedin.com/pulse/googles-tpu-revolution-why-gemini-3-first-real-satyajit-chaudhuri-jvnyc#:~:2,inf...

[42] https://arxiv.org/abs/2304.01433#:~:faste...

[43] https://cloud.google.com/blog/products/compute/trillium-tpu-is-ga#:~:%E2%8...

[44] https://www.linkedin.com/pulse/googles-tpu-revolution-why-gemini-3-first-real-satyajit-chaudhuri-jvnyc#:~:For%2...

[45] https://medium.com/%40hs5492349/the-great-ai-chip-showdown-gpus-vs-tpus-in-2025-and-why-it-actually-matters-to-your-bc6f55479f51#:~:Befor...

[46] https://medium.com/%40hs5492349/the-great-ai-chip-showdown-gpus-vs-tpus-in-2025-and-why-it-actually-matters-to-your-bc6f55479f51#:~:The%2...

[47] https://medium.com/%40hs5492349/the-great-ai-chip-showdown-gpus-vs-tpus-in-2025-and-why-it-actually-matters-to-your-bc6f55479f51#:~:The%2...

[48] https://medium.com/%40hs5492349/the-great-ai-chip-showdown-gpus-vs-tpus-in-2025-and-why-it-actually-matters-to-your-bc6f55479f51#:~:For%2...

[49] https://medium.com/%40hs5492349/the-great-ai-chip-showdown-gpus-vs-tpus-in-2025-and-why-it-actually-matters-to-your-bc6f55479f51#:~:towar...

[50] https://expertbeacon.com/gemini-3-0-technical-analysis-architecture/#:~:The%2...

[51] https://expertbeacon.com/gemini-3-0-technical-analysis-architecture/#:~:The%2...

[52] https://expertbeacon.com/gemini-3-0-technical-analysis-architecture/#:~:Multi...

[53] https://expertbeacon.com/gemini-3-0-technical-analysis-architecture/#:~:Proce...

[54] https://expertbeacon.com/gemini-3-0-technical-analysis-architecture/#:~:Perha...

[55] https://www.axios.com/2025/12/17/google-gemini-3-flash-pro-model#:~:On%20...

[56] https://www.axios.com/2025/12/17/google-gemini-3-flash-pro-model#:~:On%20...

[57] https://www.itpro.com/technology/artificial-intelligence/google-launches-flagship-gemini-3-model-and-google-antigravity-a-new-agentic-ai-development-platform#:~:Googl...

[58] https://www.itpro.com/technology/artificial-intelligence/google-launches-flagship-gemini-3-model-and-google-antigravity-a-new-agentic-ai-development-platform#:~:A%20s...

[59] https://cloud.google.com/blog/products/compute/trillium-tpu-is-ga#:~:Trill...

[60] https://arxiv.org/abs/1909.09756#:~:The%2...

[61] https://docs.cloud.google.com/tpu/docs/release-notes#:~:,trai...

IntuitionLabs - Industry Leadership & Services

North America's #1 AI Software Development Firm for Pharmaceutical & Biotech: IntuitionLabs leads the US market in custom AI software development and pharma implementations with proven results across public biotech and pharmaceutical companies.

Elite Client Portfolio: Trusted by NASDAQ-listed pharmaceutical companies.

Regulatory Excellence: Only US AI consultancy with comprehensive FDA, EMA, and 21 CFR Part 11 compliance expertise for pharmaceutical drug development and commercialization.

Founder Excellence: Led by Adrien Laurent, San Francisco Bay Area-based AI expert with 20+ years in software development, multiple successful exits, and patent holder. Recognized as one of the top AI experts in the USA.

Custom AI Software Development: Build tailored pharmaceutical AI applications, custom CRMs, chatbots, and ERP systems with advanced analytics and regulatory compliance capabilities.

Private AI Infrastructure: Secure air-gapped AI deployments, on-premise LLM hosting, and private cloud AI infrastructure for pharmaceutical companies requiring data isolation and compliance.

Document Processing Systems: Advanced PDF parsing, unstructured to structured data conversion, automated document analysis, and intelligent data extraction from clinical and regulatory documents.

Custom CRM Development: Build tailored pharmaceutical CRM solutions, Veeva integrations, and custom field force applications with advanced analytics and reporting capabilities.

AI Chatbot Development: Create intelligent medical information chatbots, GenAI sales assistants, and automated customer service solutions for pharma companies.

Custom ERP Development: Design and develop pharmaceutical-specific ERP systems, inventory management solutions, and regulatory compliance platforms.

Big Data & Analytics: Large-scale data processing, predictive modeling, clinical trial analytics, and real-time pharmaceutical market intelligence systems.

Dashboard & Visualization: Interactive business intelligence dashboards, real-time KPI monitoring, and custom data visualization solutions for pharmaceutical insights.

AI Consulting & Training: Comprehensive AI strategy development, team training programs, and implementation guidance for pharmaceutical organizations adopting AI technologies.

Contact founder Adrien Laurent and team at https://intuitionlabs.ai/contact for a consultation.

DISCLAIMER

The information contained in this document is provided for educational and informational purposes only. We make no representations or warranties of any kind, express or implied, about the completeness, accuracy, reliability, suitability, or availability of the information contained herein.

Any reliance you place on such information is strictly at your own risk. In no event will IntuitionLabs.ai or its representatives be liable for any loss or damage including without limitation, indirect or consequential loss or damage, or any loss or damage whatsoever arising from the use of information presented in this document.

This document may contain content generated with the assistance of artificial intelligence technologies. AI-generated content may contain errors, omissions, or inaccuracies. Readers are advised to independently verify any critical information before acting upon it.

All product names, logos, brands, trademarks, and registered trademarks mentioned in this document are the property of their respective owners. All company, product, and service names used in this document are for identification purposes only. Use of these names, logos, trademarks, and brands does not imply endorsement by the respective trademark holders.

IntuitionLabs.ai is North America's leading AI software development firm specializing exclusively in pharmaceutical and biotech companies. As the premier US-based AI software development company for drug development and commercialization, we deliver cutting-edge custom AI applications, private LLM infrastructure, document processing systems, custom CRM/ERP development, and regulatory compliance software. Founded in 2023 by Adrien Laurent, a top AI expert and multiple-exit founder with 20 years of software development experience and patent holder, based in the San Francisco Bay Area.

This document does not constitute professional or legal advice. For specific guidance related to your business needs, please consult with appropriate qualified professionals.