

Google announces TPU v5p with 8,960-chip pods and says Gemini was trained on TPUs

Source copy kept by wheresthe.ai for the story "Google announces TPU v5p with 8,960-chip pods and says Gemini was trained on TPUs" (<https://wheresthe.ai/d/google-tpu-v5p-ai-hypercomputer-2023-12/>).

Source: <https://cloud.google.com/blog/products/ai-machine-learning/introducing-cloud-tpu-v5p-and-ai-hypercomputer>

Text saved 18 Oct 2026 (the page as published).

SHA-256 of the saved text: 0cc01ce40d56061f801c8239f7952f689ccb7204b6f5c1d7f020a9b0677f4912

AI & Machine Learning

Enabling next-generation AI workloads: Announcing TPU v5p and AI Hypercomputer

December 6, 2023

Amin Vahdat VP/GM, Machine Learning, Systems, and Cloud AI, Google Cloud

Mark Lohmeyer VP & GM, AI & Computing Infrastructure

Generative AI (gen AI) models are rapidly evolving, offering unparalleled sophistication and capability. This advancement empowers enterprises and developers across various industries to solve complex problems and unlock new opportunities. However, the growth in gen AI models — with a tenfold increase in parameters annually over the past five years — brings heightened requirements for training, tuning, and inference. Today's larger models, featuring hundreds of billions or even trillions of parameters, require extensive training periods, sometimes spanning months, even on the most specialized systems. Additionally, efficient AI workload management necessitates a coherently integrated AI stack consisting of optimized compute, storage, networking, software and development frameworks. Today, to address these challenges, we are excited to announce Cloud TPU v5p, our most powerful, scalable, and flexible AI accelerator thus far. TPUs have long been the basis for training and serving AI-powered products like YouTube, Gmail, Google Maps, Google Play, and Android. In fact, Gemini, Google's most capable and general AI model announced today, was trained on, and is served, using TPUs. In addition, we are also announcing AI Hypercomputer from Google Cloud, a groundbreaking supercomputer architecture that employs an integrated system of performance-optimized hardware, open software, leading ML frameworks, and flexible consumption models. Traditional methods often tackle demanding AI workloads through piecemeal, component-level enhancements, which can lead to inefficiencies and bottlenecks. In contrast, AI Hypercomputer employs systems-level codesign to boost efficiency and productivity across AI training, tuning, and serving.

Inside Cloud TPU v5p, our most powerful and scalable TPU accelerator to date. Earlier this year, we announced the general availability of Cloud TPU v5e. With 2.3X price performance improvements over the previous generation TPU v4, it is our most cost-efficient TPU to date. By contrast, Cloud TPU v5p, is our most powerful TPU thus far. Each TPU v5p pod composes together 8,960 chips over our highest-bandwidth inter-chip interconnect (ICI) at 4,800 Gbps/chip in a 3D torus topology. Compared to TPU v4, TPU v5p features more than 2X greater FLOPS and 3X more high-bandwidth memory (HBM). Designed for performance, flexibility, and scale, TPU v5p can train large LLM models 2.8X faster than the previous-generation TPU v4. Moreover, with second-generation SparseCores, TPU v5p can train embedding-dense models 1.9X faster than TPU v4.

Source: TPU v5p and v4 are based on Google Internal Data. As of November, 2023: All numbers normalized per chip seq-len=2048 for GPT-3 175 billion parameter model.

Source: TPU v5e data is from MLPerf™ 3.1 Training Closed results for v5e. TPU v5p and v4 are based on Google internal training runs. As of November, 2023: All numbers normalized per chip seq-len=2048 for GPT-3 175 billion parameter model. It shows relative performance per dollar using the public list price of TPU v4 (\$3.22/chip/hour), TPU v5e (\$1.2/chip/hour) and TPU v5p (\$4.2/chip/hour).

In addition to performance improvements, TPU v5p is also 4X more scalable than TPU v4 in terms of total available FLOPs per pod. Doubling the floating-point operations per second (FLOPS) over TPU v4 and doubling the number of chips in a single pod provides considerable improvement in relative performance in training speed.

Google AI Hypercomputer delivers peak performance and efficiency at large scale. Achieving both scale and

speed is necessary, but not sufficient to meet the needs of modern AI/ML applications and services. The hardware and software components must come together into an integrated, easy-to-use, secure, and reliable computing system. At Google, we've done decades of research and development on this very problem, culminating in AI Hypercomputer, a system of technologies optimized to work in concert to enable modern AI workloads.

Performance-optimized hardware: AI Hypercomputer features performance-optimized compute, storage, and networking built over an ultrascale data center infrastructure, leveraging a high-density footprint, liquid cooling, and our Jupiter data center network technology. All of this is predicated on technologies that are built with efficiency at their core; leveraging clean energy and a deep commitment to water stewardship, and that are helping us move toward a carbon-free future. **Open software:** AI Hypercomputer enables developers to access our performance-optimized hardware through the use of open software to tune, manage, and dynamically orchestrate AI training and inference workloads on top of performance-optimized AI hardware. Extensive support for popular ML frameworks such as JAX, TensorFlow, and PyTorch are available right out of the box. Both JAX and PyTorch are powered by OpenXLA compiler for building sophisticated LLMs. XLA serves as a foundational backbone, enabling the creation of complex multi-layered models (Llama 2 training and inference on Cloud TPUs with PyTorch/XLA). It optimizes distributed architectures across a wide range of hardware platforms, ensuring easy-to-use and efficient model development for diverse AI use cases (AssemblyAI leverages JAX/XLA and Cloud TPUs for large-scale AI speech). Open and unique Multislice Training and Multihost Inferencing software, respectively, make scaling, training, and serving workloads smooth and easy. Developers can scale to tens of thousands of chips to support demanding AI workloads. Deep integration with Google Kubernetes Engine (GKE) and Google Compute Engine, to deliver efficient resource management, consistent ops environments, autoscaling, node-pool auto-provisioning, auto-checkpointing, auto-resumption, and timely failure recovery.

Flexible consumption : AI Hypercomputer offers a wide range of flexible and dynamic consumption choices. In addition to classic options, such as Committed Use Discounts (CUD), on-demand pricing, and spot pricing, AI Hypercomputer provides consumption models tailored for AI workloads via Dynamic Workload Scheduler. Dynamic Workload Scheduler introduces two models: Flex Start mode for higher resource obtainability and optimized economics, as well as Calendar mode, which targets workloads with higher predictability on job-start times. Leveraging Google's deep experience to help power the future of AI Customers like Salesforce and Lightricks are already training and serving large AI models with Google Cloud's TPU v5p AI Hypercomputer — and already seeing a difference: "We've been leveraging Google Cloud TPU v5p for pre-training Salesforce's foundational models that will serve as the core engine for specialized production use cases, and we're seeing considerable improvements in our training speed. In fact, Cloud TPU v5p compute outperforms the previous generation TPU v4 by as much as 2X. We also love how seamless and easy the transition has been from Cloud TPU v4 to v5p using JAX. We're excited to take these speed gains even further by leveraging the native support for INT8 precision format via the Accurate Quantized Training (AQT) library to optimize our models." - Erik Nijkamp, Senior Research Scientist, Salesforce "Leveraging the remarkable performance and ample memory capacity of Google Cloud TPU v5p, we successfully trained our generative text-to-video model without splitting it into separate processes. This optimal hardware utilization significantly accelerates each training cycle, allowing us to swiftly conduct a series of experiments. The ability to train our model quickly in each experiment facilitates rapid iteration, which is an invaluable advantage for our research team in this competitive field of generative AI." - Yoav HaCohen, PhD, Core Generative AI Research Team Lead, Lightricks "In our early-stage usage, Google DeepMind and Google Research have observed 2X speedups for LLM training workloads using TPU v5p chips compared to the performance on our TPU v4 generation. The robust support for ML Frameworks (JAX, PyTorch, TensorFlow) and orchestration tools enables us to scale even more efficiently on v5p. With the 2nd generation of SparseCores we also see significant improvement in the performance of embeddings-heavy workloads. TPUs are vital to enabling our largest-scale research and engineering efforts on cutting edge models like Gemini." - Jeff Dean, Chief Scientist, Google DeepMind and Google Research At Google, we've long believed in the power of AI to help solve challenging problems. Until very recently, training large foundation models and serving them at scale was too complicated and expensive for many organizations. Today, with Cloud TPU v5p and AI Hypercomputer, we're excited to extend the result of decades of research in AI and systems design with our customers, so they can innovate with AI faster, more efficiently, and more cost effectively. To request access to Cloud TPU v5p and AI Hypercomputer, please reach out to your Google Cloud account manager . 1: MLPerf™ v3.1 Training Closed, multiple benchmarks as shown. Retrieved November 8th, 2023 from mlcommons.org . Results 3.1-2004. Performance per dollar is not an MLPerf metric. TPU v4 results are unverified: not verified by MLCommons Association. The MLPerf™ name and logo are trademarks of MLCommons Association in the United States and other countries. All rights reserved. Unauthorized use strictly prohibited. See www.mlcommons.org for more

information. 2: Google Internal Data for TPU v5p as of November, 2023: E2E steptime, SearchAds pCTR, batch size per TPU core 16,384, 125 v5p chips

Posted in AI & Machine Learning Compute Infrastructure Modernization Systems

Related articles

Healthcare & Life Sciences Architecting autonomy: A design pattern for high-compute, governed agents in biotech
By Schneider Larbi • 9-minute read

Developers & Practitioners Four evaluation engineering lessons learned from building agent plugins By Casey West • 5-minute read

Small and Medium Businesses What the Gemini agent means for startups By Darren Mowry • 6-minute read

Media & Entertainment How Visko streams real-time world models on Google Cloud TPUs By Will Yin • 8-minute read