Page Contents

Announcing the general availability of **Trillium**, our sixth generation and most advanced TPU to date.

Cloud Tensor Processing Units (TPUs)

Accelerate AI development with Google Cloud TPUs

Cloud TPUs optimize performance and cost for all AI workloads, from training to inference. Using world-class data center infrastructure, TPUs offer high reliability, availability, and security.

Get started

Not sure if TPUs are the right fit? Learn about when to use GPUs or CPUs on Compute Engine instances to run your machine learning workloads.

Product highlights

- Run large-scale AI training workloads
- Fine-tune foundational AI models
- Serve large-scale AI inference workloads

Next Generation Infrastructure Innovations for Large-Scale AI Models

OVERVIEW

What is a Tensor Processing Unit (TPU)?

Google Cloud TPUs are custom-designed AI accelerators, which are optimized for training and inference of large AI models. They are ideal for a variety of use cases, such as chatbots, code generation, media content generation, synthetic speech, vision services, recommendation engines, personalization models, among others.

What are the advantages of Cloud TPUs?

Cloud TPUs are designed to scale cost-efficiently for a wide range of AI workloads, spanning training, fine-tuning, and inference. Cloud TPUs provide the versatility to accelerate workloads on leading AI frameworks, including PyTorch, JAX, and TensorFlow. Seamlessly orchestrate large-scale AI workloads through Cloud TPU integration in Google Kubernetes Engine (GKE). Leverage Dynamic Workload Scheduler to improve the scalability of workloads by scheduling all accelerators needed simultaneously. Customers looking for the simplest way to develop AI models can also leverage Cloud TPUs in Vertex AI, a fully-managed AI platform.

When to use Cloud TPUs?

Cloud TPUs are optimized for training large and complex deep learning models that feature many matrix calculations, for instance building large language models (LLMs). Cloud TPUs also have SparseCores, which are dataflow processors that accelerate models relying on embeddings found in recommendation models. Other use cases include healthcare, like protein folding modeling and drug discovery.

How are Cloud TPUs different from GPUs?

A GPU is a specialized processor originally designed for manipulating computer graphics. Their parallel structure makes them ideal for algorithms that process large blocks of data commonly found in AI workloads. Learn more.

A TPU is an application-specific integrated circuit (ASIC) designed by Google for neural networks. TPUs possess specialized features, such as the matrix multiply unit (MXU) and proprietary interconnect topology that make them ideal for accelerating AI training and inference.

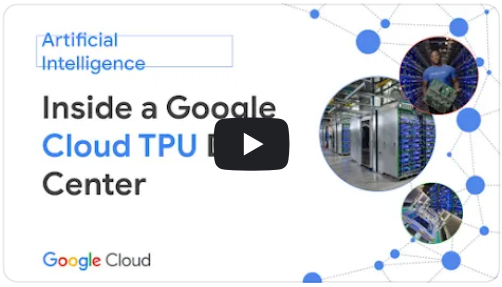
Cloud TPU versions

Cloud TPU version	Description	Availability
Trillium	The most advanced Cloud TPU to date	During preview, Trillium is available in North America (US East region), Europe (West region), and Asia (Northeast region)
Cloud TPU v5p	The most powerful Cloud TPU for training AI models	Cloud TPU v5p is generally available in North America (US East region)
Cloud TPU v5e	A versatile Cloud TPU for training and inference needs	Cloud TPU v5e is generally available in North America (US Central/East/South/ West regions), Europe (West region), and Asia (Southeast region)

Additional information on Cloud TPU versions

HOW IT WORKS

Get an inside look at the magic of Google Cloud TPUs, including a rare inside view of the data centers where it all happens. Customers use Cloud TPUs to run some of the world's largest AI workloads and that power comes from much more than just a chip. In this video, take a look at the components of the TPU system, including data center networking, optical circuit switches, water cooling systems, biometric security verification and more.



COMMON USES

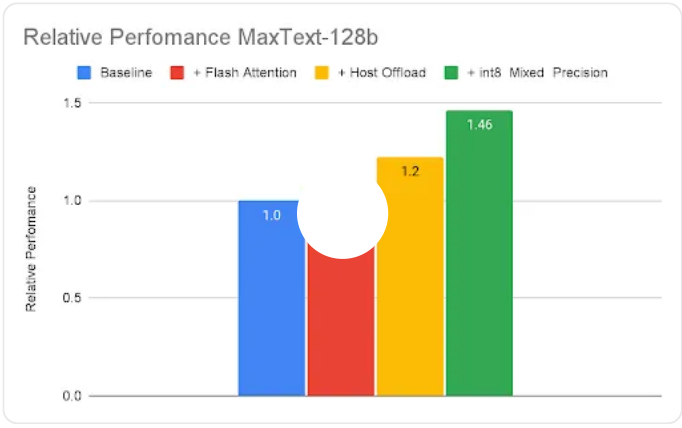
Run large-scale AI training workloads

How-tos | Additional resources

Performant and efficient model training

Get started quickly with MaxText and MaxDiffusion, high performance, highly scalable open source reference deployments for large model training.

Learn more



Train MaxText using a synthetic dataset on Cloud TPU

Train Llama 2 with PyTorch/XLA on TPU v5p

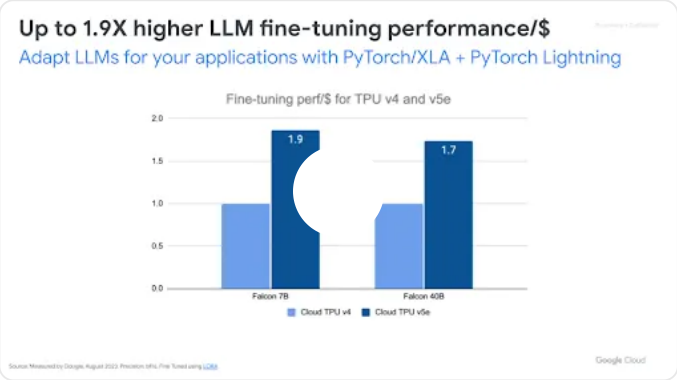
Train Hugging Face FLAX models on TPU v5e

Fine-tune foundational AI models

Additional resources

Adapt LLMs for your applications with Pytorch/XLA

Efficiently fine-tune foundation models by leveraging your own training data that represents your use case. Cloud TPU v5e provides up to 1.9x higher LLM fine-tuning performance per dollar compared to Cloud TPU v4.



Serve large-scale AI inference workloads

How-tos | Additional resources

High-performance, scalable, cost-efficient inference

Accelerate AI Inference with JetStream and MaxDiffusion. JetStream is a new inference engine specifically designed for Large Language Model (LLM) inference. JetStream represents a significant leap forward in both performance and cost efficiency, offering unparalleled throughput and latency for LLM inference on Cloud TPUs. MaxDiffusion is a set of diffusion model implementations optimized for Cloud TPUs, making it easy to run inference for diffusion models on Cloud TPUs with high performance.

Learn more

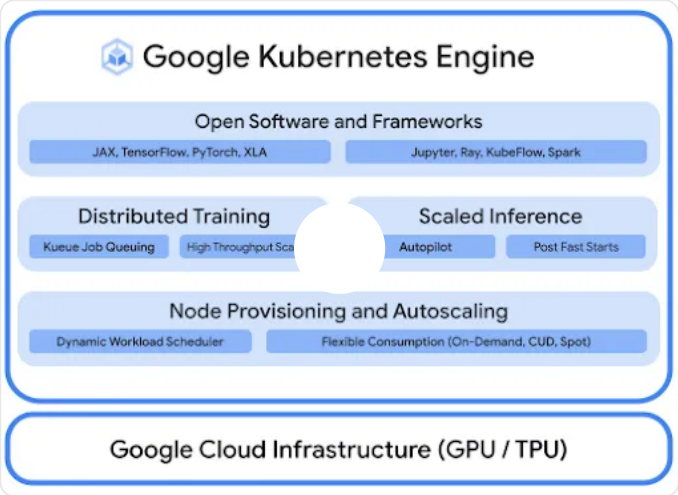
- JetStream MaxText Inference on v5e Cloud TPU VM
- JetStream PyTorch Inference on v5e Cloud TPU VM
- Deploy a HuggingFace TGI server on a Google Cloud TPU instance

Cloud TPU in GKE

How-tos | Additional resources

Run optimized AI workloads with platform orchestration

A robust AI/ML platform considers the following layers: (i) Infrastructure orchestration that support GPUs for training and serving workloads at scale, (ii) Flexible integration with distributed computing and data processing frameworks, and (iii) Support for multiple teams on the same infrastructure to maximize utilization of resources.



Learn more about AI/ML orchestration on GKE

Serve Gemma using TPUs on GKE with JetStream

Run Ray on GKE with TPUs

Cloud TPU in Vertex AI

Additional resources

Vertex AI training and predictions with Cloud TPUs

For customers looking for a simplest way to develop AI models, you can deploy Cloud TPU v5e with Vertex AI, an end-to-end platform for building AI models on fully-managed infrastructure that's purpose-built for low-latency serving and high-performance training.

Vertex AI training with Cloud TPUs

Vertex AI predictions with Cloud TPUs

PRICING

Cloud TPU pricing

All Cloud TPU pricing is per chip-hour

Cloud TPU Version	Evaluation Price (USD)	1-year commitment (USD)	3-year commitment (USD)
Trillium	Starting at **$2.7000** per chip-hour	Starting at **$1.8900** per chip-hour	Starting at **$1.2200** per chip-hour
Cloud TPU v5p	Starting at **$4.2000** per chip-hour	Starting at **$2.9400** per chip-hour	Starting at **$1.8900** per chip-hour
Cloud TPU v5e	Starting at **$1.2000** per chip-hour	Starting at **$0.8400** per chip-hour	Starting at **$0.5400** per chip-hour

Cloud TPU pricing varies by product and region.

PRICING CALCULATOR

Estimate your monthly Cloud TPU costs, including region specific pricing and fees.

Estimate your costs

CUSTOM QUOTE

Connect with our sales team to get a custom quote for your organization.

Request a quote

Start your proof of concept

Try Cloud TPUs for free

Get started

Get a quick intro to using Cloud TPUs

Learn more

Run TensorFlow on Cloud TPU VM

View guide

Run JAX on Cloud TPU VM

View guide

Run PyTorch on Cloud TPU VM

View guide