

Cloud TPU

Train and run machine learning models faster than ever before.

[View documentation \(https://cloud.google.com/tpu/docs/\)](https://cloud.google.com/tpu/docs/)

[Get started \(https://cloud.google.com/tpu/docs/quickstart\)](https://cloud.google.com/tpu/docs/quickstart)



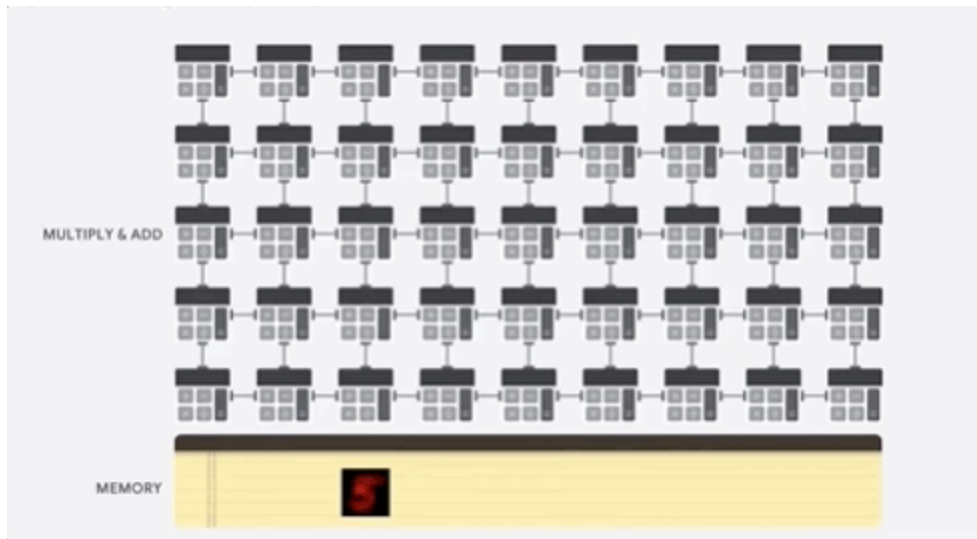
Empowering businesses with Google Cloud AI

Machine learning has produced business and research breakthroughs ranging from network security to medical diagnoses. We built the Tensor Processing Unit (TPU) in order to make it possible for anyone to achieve similar breakthroughs. Cloud TPU is the custom-designed machine learning ASIC that powers Google products like Translate, Photos, Search, Assistant, and Gmail. Here's how you can put the TPU and machine learning to work accelerating your company's success, especially at scale.

Try the demo

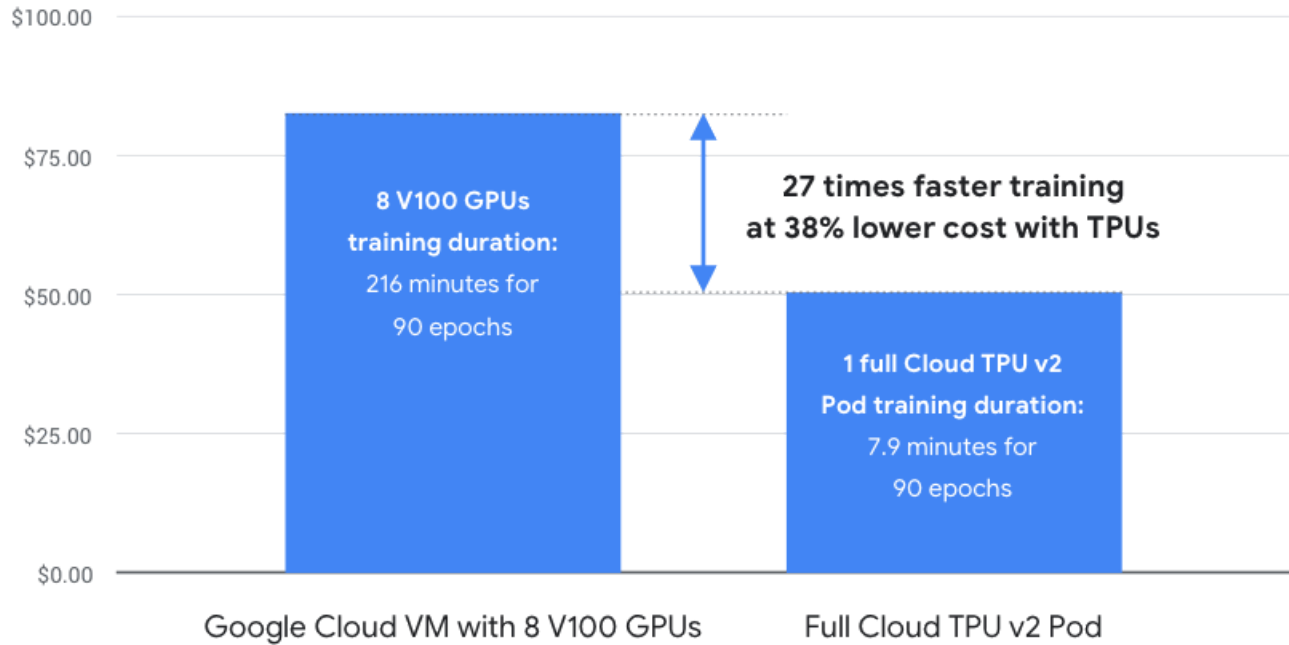
Want to learn more about how the Cloud TPU is fine-tuned for your machine learning applications?

Give the [demo \(//tpudemo.com\)](https://tpudemo.com) a try or read the [blog \(https://cloud.google.com/blog/products/ai-machine-learning/what-makes-tpus-fine-tuned-for-deep-learning\)](https://cloud.google.com/blog/products/ai-machine-learning/what-makes-tpus-fine-tuned-for-deep-learning)



Machine learning performance and benchmarks

ResNet-50 Training Cost Comparison



To see how Cloud TPU compares to other accelerators, read the [blog](https://cloud.google.com/blog/products/ai-machine-learning/cloud-tpu-pods-break-ai-training-records) (https://cloud.google.com/blog/products/ai-machine-learning/cloud-tpu-pods-break-ai-training-records)

or view the MLPerf benchmark [results](https://mlperf.org/training-results-0-6/) (//mlperf.org/training-results-0-6/).

Benefits



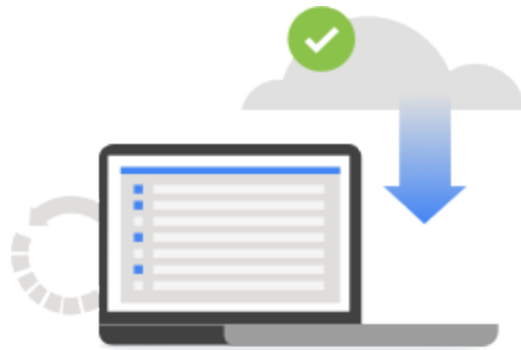
Built for AI on Google Cloud

Cloud TPU is designed to run cutting-edge machine learning models with AI services on Google Cloud. And its custom high-speed network offers over 100 petaflops of performance in a single pod – enough computational power to transform your business or create the next research breakthrough.



Iterate faster on your ML solutions

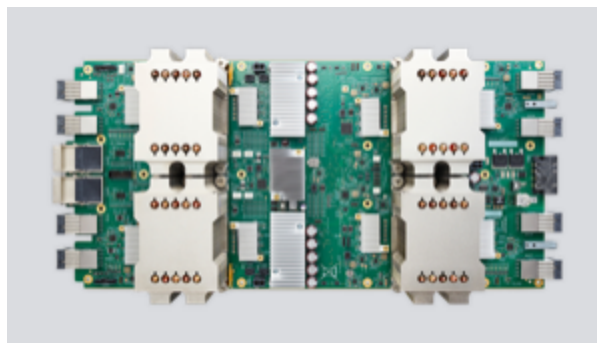
Training machine learning models is like compiling code: you need to update often, and you want to do so as efficiently as possible. ML models need to be trained over and over as apps are built, deployed, and refined. Cloud TPU's robust performance and low cost make it ideal for machine learning teams looking to iterate quickly and frequently on their solutions.



Proven, state-of-the-art models

You can build your own machine learning-powered solutions for many real-world use cases. Just bring your data, download a Google-optimized reference model, and start training.

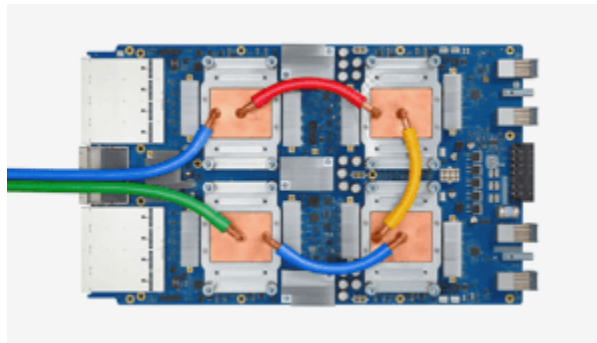
Cloud TPU offering



Cloud TPU v2

180 teraflops

64 GB High Bandwidth Memory (HBM)



Cloud TPU v3

420 teraflops

128 GB HBM

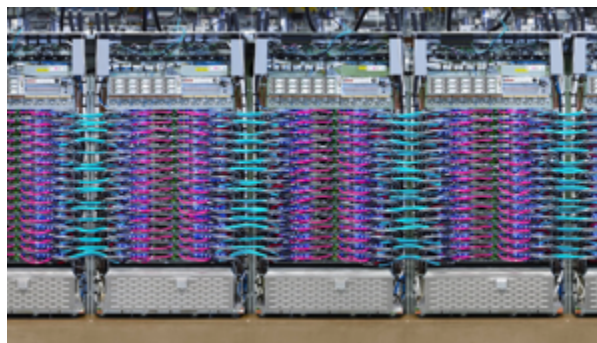


Cloud TPU v2 Pod

11.5 petaflops

4 TB HBM

2-D toroidal mesh network



Cloud TPU v3 Pod

100+ petaflops

32 TB HBM

2-D toroidal mesh network

Cloud TPU features

Model library

Get started immediately by leveraging our growing [library of optimized models](https://github.com/tensorflow/tpu/tree/master/models/official) ([//github.com/tensorflow/tpu/tree/master/models/official](https://github.com/tensorflow/tpu/tree/master/models/official)) for Cloud TPU. These provide optimized performance, accuracy, and quality in image classification, object detection, language modeling, speech recognition, and more.

Connect Cloud TPUs to custom machine types

You can connect to Cloud TPUs from custom [AI Platform Deep Learning VM Image](https://cloud.google.com/deep-learning-vm/) (<https://cloud.google.com/deep-learning-vm/>) types, which can help you optimally balance processor speeds, memory, and high-performance storage resources for your workloads.

Fully integrated with Google Cloud Platform

At their core, Cloud TPUs and Google Cloud's [data and analytics](https://cloud.google.com/products/big-data/) (<https://cloud.google.com/products/big-data/>) services are fully integrated with other Google Cloud Platform offerings, like [Google Kubernetes Engine](https://cloud.google.com/kubernetes-engine/) (<https://cloud.google.com/kubernetes-engine/>) (GKE). So when you run machine learning workloads on Cloud TPUs, you benefit from GCP's industry-leading [storage](#)

(<https://cloud.google.com/storage/>), [networking](#) (<https://cloud.google.com/products/networking/>), and [data analytics](#) (<https://cloud.google.com/products/big-data/>) technologies.

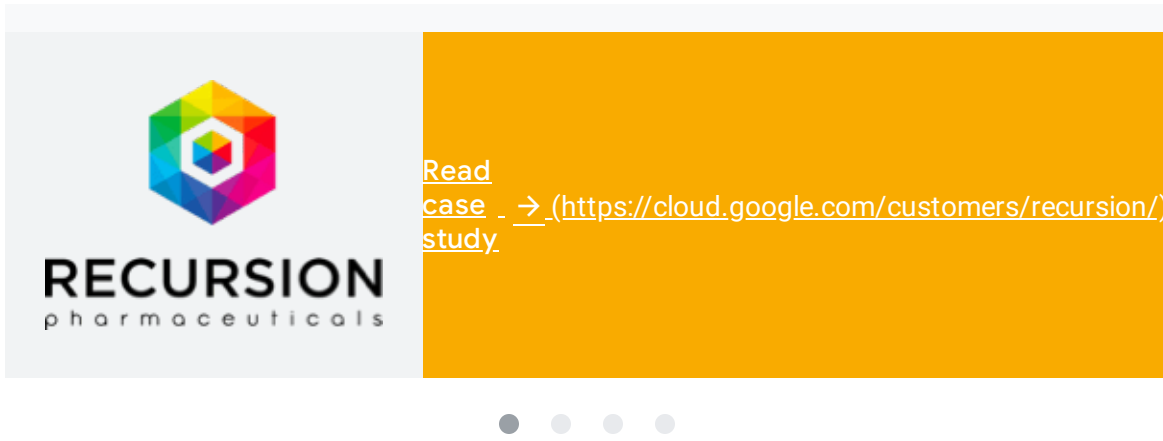
Preemptible Cloud TPU

You can save money by using preemptible Cloud TPUs for fault-tolerant machine learning workloads, such as long training runs with checkpointing or batch prediction on large datasets. Preemptible Cloud TPUs are 70% cheaper than on-demand instances, making everything from your first experiments to large-scale hyperparameter searches more affordable than ever.

Our customers

“ The potential of using Cloud TPU pods to accelerate our deep learning research while keeping operational costs and complexity low is a big draw. It takes us now a little over 24 hours to train models on our local GPU cluster. It will take us, depending on the size of the TPU pod, anywhere from 7 hours to 15 minutes. ”

Ben Mabey, Vice President, Engineering, Recursion Pharmaceuticals



[See all customers → \(https://cloud.google.com/customers/\)](https://cloud.google.com/customers/)

Pricing

Cloud TPU charges for using preemptible and non-preemptible (on-demand) Cloud TPU use to train machine learning models. For detailed pricing information, please view the [pricing guide \(https://cloud.google.com/tpu/docs/pricing\)](https://cloud.google.com/tpu/docs/pricing).

Single Cloud TPU device pricing

The following table shows the pricing per region for using a single Cloud TPU device.

US EUROPE ASIA PACIFIC

Version	On-demand	Preemptible
Cloud TPU v2	\$4.50 / TPU hour	\$1.35 / TPU hour
Cloud TPU v3	\$8.00 / TPU hour	\$2.40 / TPU hour

Cloud TPU Pod pricing

The following table shows the pricing for using a Cloud TPU Pod slices.

--

Cloud TPU v2 Pod	Evaluation Price / hr	1-yr Commitment Price (37% discount)	3-yr Commitment Price (55% discount)
32-core Pod slice	\$24 USD	\$132,451 USD	\$283,824 USD
128-core Pod slice	\$96 USD	\$529,805 USD	\$1,135,296 USD
256-core Pod slice	\$192 USD	\$1,059,610 USD	\$2,270,592 USD
512-core Pod slice	\$384 USD	\$2,119,219 USD	\$4,541,184 USD
Cloud TPU v3 Pod	Evaluation Price / hr	1-yr Commitment Price (37% discount)	3-yr Commitment Price (55% discount)
32-core Pod slice	\$32 USD	\$176,601 USD	\$378,432 USD

To request a Cloud TPU Pod configuration or a quote for larger Cloud TPU v3 Pod slices, please [contact a sales representative](https://cloud.google.com/contact/) (<https://cloud.google.com/contact/>).


Resources



Cloud TPU tutorials, quickstarts, and docs

View documentation → (<https://cloud.google.com/tpu/docs>)

Next '19: Fast and lean data science with TPUs

Watch video  (<https://www.youtube.com/watch?v=mvYqa0I2g68&feature=youtu.be>)



Cloud TPU content from AI Hub

Learn more ↗ (<https://aihub.cloud.google.com/publications/530-4899-a4e2-1e02b48ffb1b>)



What's in an image: fast, accurate image segmentation with Cloud TPUs

Read blog → (<https://cloud.google.com/blog/topics/machine-learning/whats-in-an-image-accurate-image-segmentation-with-tpus>)



What makes TPUs fine-tuned for deep learning?

Read blog → (<https://cloud.google.com/blog/topics/machine-learning/what-makes-tpus-fine-tuned-for-deep-learning>)



Train TensorFlow ML models faster and at lower cost on Cloud TPU Pods

Read blog → (<https://cloud.google.com/blog/topics/machine-learning/now-you-cant-train-tensorflow-ml-models-faster-and-lower-cost-on-cloud-tpu-pods>)

Cloud AI products comply with the SLA policies listed [here](https://cloud.google.com/terms/sla/) (https://cloud.google.com/terms/sla/). They may offer different latency or availability guarantees from other Google Cloud services.