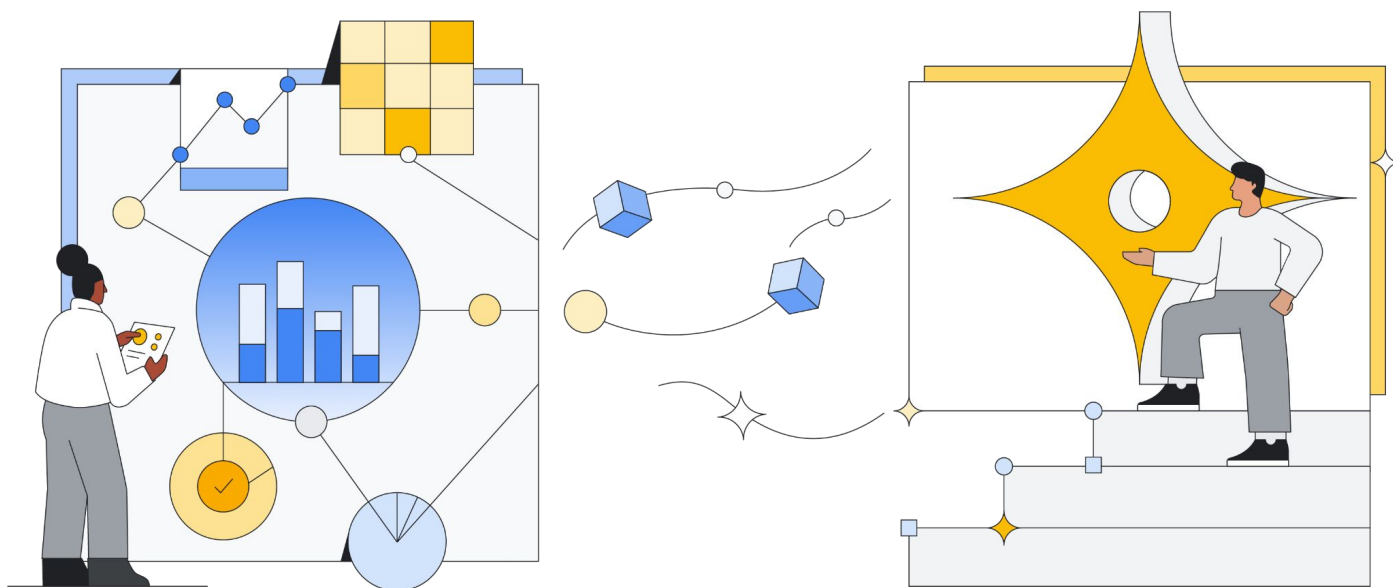


The future of AI starts with infrastructure



Your deep-dive guide to
Google Cloud AI infrastructure

Table of contents

Introduction

- 3 AI is powering the future of business
- 4 What is AI infrastructure?
- 5 Top trends driving the transformation of AI and AI infrastructure
- 8 Why Google is the leader in AI
- 9 How to unlock more value from your AI
- 11 How to know which AI infrastructure is right for you

Solution deep dive

- 13 Optimize performance and costs with fully-managed AI infrastructure
- 17 Support diverse workloads with flexible and open infrastructure
- 25 Innovate faster with state-of-the-art AI solutions
- 28 Next steps



AI is powering the future of business

“Can machines think?” The world has come a long way since Alan Turing famously posed this question in 1950. Today, AI is applied across industries and throughout our lives—helping us with everything from choosing what songs to play, to finding urgent solutions to global health problems.

Indeed, AI has progressed to the point where it now augments human understanding; and machines can mimic human behavior to complete complex tasks in real time. With endless opportunities ahead of them, companies are racing to adopt AI and build exciting new applications that drive their business forward.

Yet, as this field of computing grows and as access to data increases, AI models are getting larger. The cost to train, deploy, and manage these models is adding up—and it can become prohibitive for companies to take their next steps. As companies move from exploring and experimenting with ML to deploying applications at scale, they need the right AI infrastructure to keep their models performant, while staying cost effective.

So, what is AI infrastructure?



Most people think of hardware when they think of AI infrastructure. In reality, a combination of three critical ingredients work together to enable companies to access and manage the resources needed to train and deploy advanced AI workloads.

First, there's the high-performing, cost-efficient **compute**, which is engineered to enable businesses to deliver AI at scale.

Then there are the **ML platforms and tools** which help teams access and efficiently manage the resources needed to train and deploy AI workloads. Finally, **open source software** (OSS) and solutions help teams get the most out of their underlying infrastructure by making it faster and easier to innovate and scale ML across an enterprise.

3 key ingredients in Google Cloud AI infrastructure



Open source software

Open source 'State of the Art (SOTA) AI' solutions

Open source modeling frameworks fully integrated with Google Cloud (Tensorflow, PyTorch)

Support for open source custom modeling (e.g. PyTorch, TensorFlow) and open source opinionated SOTA (e.g. AlphaFold, TabNet, Merlin)



ML Platforms and tools

Vertex AI
fully-managed ML platform

Building ML-specific services with Vertex AI for notebooks, training, serving and so on



Hardware

Purpose-built AI compute and accelerators

GPUs, TPU, CPUs, networking

Forrester names Google Cloud a leader in AI infrastructure based on its AI architecture, training, inference, and management capabilities.¹

1. [Forrester](#)

Top trends driving the transformation of AI and AI infrastructure

Companies are constantly on the lookout for new ways to use AI to delight customers, streamline operations, and grow their business. And, as AI adoption skyrockets, rapid advancements in AI infrastructure are supporting the massive amounts of data and real-time inference demanded of these intelligent applications.

There are three key trends driving both the exponential growth of AI and the evolution of AI infrastructure.

1 Powerful personalization of the customer experience

A company's biggest differentiator is the way it interacts with end users. And, with customer expectations sky-high these days, companies need to think laterally about how to stand out from the crowd. Leaders are leveraging data and AI to take personalization to the next level.

For example, **AI-enabled recommendation systems** analyze large-scale data to deliver compelling, meaningful content to customers from a large collection of items—even showing items that the customer might not have considered previously. To deliver these accurate and ultimately appealing recommendations fast, companies need exceptional AI infrastructure behind them.



Wayfair delivers personalized recommendations to 30M customers

With 30 million active customers on its website, Wayfair knew that sales would grow if they could get the perfect product in front of customers early in their shopping journey. Yet their homegrown models and tools were holding them back—they lacked the AI infrastructure to deliver real-time and intelligent recommendations at scale.

Since adopting Vertex AI, Wayfair observed 5-10X faster training runs with tens of millions of data sets distributed across clusters. And, with little code written to achieve the scale, data scientists have more trust over their ecosystem and can deliver amazing, highly scalable solutions built on top of Vertex AI. They no longer need to spend effort solving complex infrastructure challenges as it abstracts those complexities away.



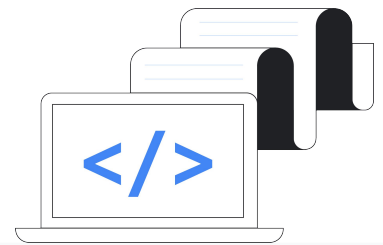
[Read Wayfair's story](#)

2

A deeper dive into knowledge extraction and understanding

AI has the capacity to translate, summarize, extract, and query large volumes of information in order to make it easier for humans to understand. And, with language the essence of how we communicate, connect, and understand the world, **Large Language Models (LLMs)** have emerged as a powerful tool in the AI landscape.

As well as the ability to predict future words in a sentence—similar to how humans talk and write—the rich understanding of language in LLMs helps with everything from code completion and genome sequencing, to human-like interactions in customer service. For companies to adopt LLMs, they need the right infrastructure and tools to support petabytes of text data and billions of parameters.



AI21labs

AI21 Labs rewrites the rules on natural language processing

To give machines the ability to write and understand natural text, AI21 Labs has built a natural language processing (NLP) model with a staggering 178 billion data parameters. A model of this scale requires high-performance GPUs and cutting edge ML capabilities—which AI21 Labs gets with Google Cloud.

Using Google Kubernetes Engine, AI21 Labs can adjust infrastructure to meet demand in a cost-effective way. And, when it comes to optimizing its NLP model, they can create multiple, identical production environments that scale based on their needs. Since adopting Google Cloud, AI21 Labs has realized huge performance gains and cost savings—with a **20% reduction on costs with no performance trade-off.**



[Read AI21 Lab's story](#)

3

Streamlining operations through automation

AI helps companies run their operations more efficiently, saving time and reducing operations costs. With the ability to make autonomous decisions in real time, ML models are powering use cases such as autonomous vehicles, factory automation, forecasting, defect detection, and more. These use cases rely on large amounts of text, image, and video data that must be analyzed in real time—leading to growing demand for purpose-built infrastructure and tools with ultra-low latency that can scale in production without a decrease in performance.



nuro

Nuro build the future of autonomous robotics

Robotics company Nuro develops autonomous last mile delivery vehicles. To help these vehicles make sense of the world around them, Nuro’s data scientists use Google Cloud’s AI infrastructure to build, train and deploy computer vision models that interpret massive amounts of images per second and react in real time.

More training data produces more accurate models, which is critically important for autonomous vehicles to interact safely in the real world. Nuro collects 100s of petabytes of time series data per week and uses the data for training and inference. Google Cloud’s flexible storage options and wide variety of high performance GPUs make it possible for Nuro to manage costs and maintain low-latency inference as their AI workloads become larger.











[Hear from Albert Meixner, Head of Software at Nuro](#)

Why Google is the leader in AI

At Google, AI is in our DNA. For decades, we have run our own scaled, diverse ML workloads on our AI infrastructure, products, and solutions—continually optimizing for performance, scalability, and cost. Now, it’s available to customers through Google Cloud’s AI infrastructure offering.

Google leverages AI to solve some of today’s toughest challenges

 <p>Search Search ranking, Speech recognition</p>	 <p>Translate Text, graphic and speech translations</p>	 <p>Photos Photos search</p>	 <p>Gmail Smart reply, Spam classification</p>
 <p>Self Driving Car 20M miles of autonomous driving</p>	 <p>Data Center Power Usage Reduced cooling energy 40%</p>	 <p>AlphaFold Creating a database of over 200M catalogued protein structures</p>	 <p>YouTube Video recommendations, Better thumbnails</p>

Did you know?



Google’s ML mega cluster (MLMC) in Oklahoma—powered by Cloud TPU v4 with nine exaflops of compute—is the **world’s largest ML hub**, operating at or near 90% clean energy. MLMC will enable developers to efficiently train the largest and most complex NLP workloads, recommendation systems, and computer vision algorithms.

**Up to
5x faster**

To get models to production with Vertex AI’s fully-managed ML platform

**Up to
8x higher**

Throughput and 6.5x lower latency with Vertex AI optimized TensorFlow Runtime for tabular data²

**Up to
40% less**

Training time and cost with Vertex AI Reduction Server and NVIDIA GPUs for large language models³

**Up to
80% faster**

With Google Cloud’s TPUv4 for large scale training workloads⁴

#1

Google is the leading Cloud Native Computing Foundation (CNCF) contributor

2. [Google](#)
3. [Google](#)
4. [Google](#)

How to unlock more value from AI

[According to IDC](#), many AI projects fail because of inadequate or lack of purpose-built infrastructure capabilities. The problem is, setting up and managing scalable and flexible AI infrastructure on your own can be an exceptionally difficult task. It's why so many companies turn to Google Cloud.

Google Cloud's full stack of AI infrastructure capabilities enable customers to build the right infrastructure for their needs, with tools they're already familiar with. From basic infrastructure building blocks to fully-managed MLOps tools and cutting-edge AI, Google Cloud has the strongest offering in the market today for AI architecture.

1 Optimize performance and costs with fully-managed AI infrastructure

AI workloads are incredibly diverse. They contain massive amounts of data. And they have high performance and latency requirements. All these things combine to make AI infrastructure management an exceptionally difficult task.

With managed AI infrastructure, Vertex AI gives data scientists easy access to a wide variety of compute, storage, and networking capabilities that meet their specific performance and budget needs. Teams without infrastructure expertise can

now train and serve models with higher performance and lower latency much faster and more cost-effectively than ever before. And Vertex AI's end-to-end MLOps capabilities make it easier to build, train, orchestrate, deploy, and manage ML at scale—reducing the work needed to optimize model performance in production.



[See page 14 to learn more](#)

With fully managed AI infrastructure on Vertex AI, customers achieved the following results:

508

Hours of time saved per model

397%

Return on investment

70%

Improved ML lifecycle efficiencies

Efficiencies in ML lifecycle valued at **\$4.5M**



2

Support diverse workloads with flexible and open infrastructure that meets your needs

Whatever you're building, Google Cloud supports large and diverse workloads, with the choice of GPUs, TPUs, or CPUs for everything from high performance training and low cost inference, to large-scale data processing.

At Google Cloud, we believe in giving you the ultimate freedom to pick the infrastructure or ML framework that works best for your business. As well as our managed infrastructure offering on Vertex AI, we offer a rich set of building blocks to help you architect your own custom software stack such as deep learning VMs on Google Compute Engine (GCE), deep learning containers on Google Kubernetes Engine (GKE), and marketplace of curated ISV offerings. Our AI infrastructure services are embedded with the most popular ML frameworks (TensorFlow, PyTorch, JAX, and MXNet)—so you can keep using whichever framework you prefer.

Did you know?



TensorFlow

Google Cloud is the only cloud provider of Tensor Processor Units (TPUs). Cloud TPU is designed to run cutting-edge ML models with AI services on Google Cloud. And its custom high-speed network offers over 1.1 exaflops of performance in a single pod. That's enough computational power to transform your business or deliver your next research breakthrough.



[See page 18 to learn more](#)

3

Innovate faster with state-of-the-art AI solutions

Developing the infrastructure to support today's leading-edge solutions is hard. Google Cloud is the only cloud provider to offer fully-managed, state-of-the-art AI solutions with optimized infrastructure—making it easier for companies to adopt the latest AI technology from Google and our partners.

Through Google Research, we deliver features in familiar tooling sets which are scaled up from discovery to enterprise-readiness within as little

as two years. Vertex AI makes it easier for data scientists to run these new innovations on known platforms like TFX, Jupyter Notebooks, and scikit-learn without code—so they can build, train, and deploy powerful AI models faster than ever.



[See page 26 to learn more](#)

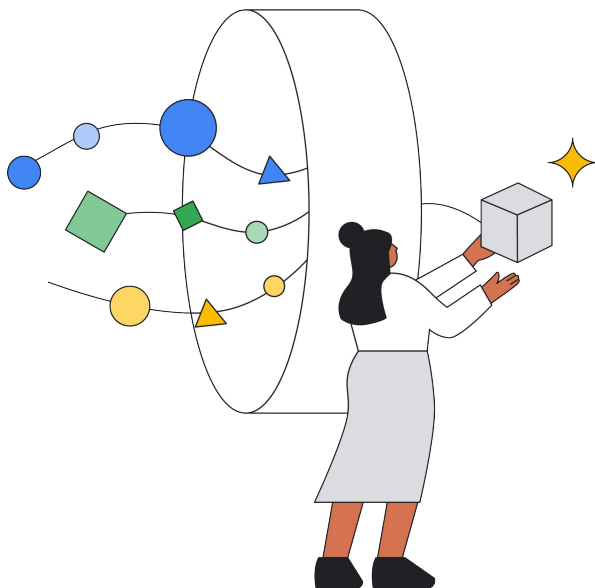
How to know which AI infrastructure is right for you

Every organization and ML project has unique requirements, and there are many ways in with AI infrastructure. Question is, which way is right for you?

With Google Cloud, you can choose our fully-managed ML platform, [Vertex AI](#); GKE and GCE to custom build with OSS and self-managed resources; or a combination of both.

Most companies—particularly those who want to build or grow ML quickly and sustainably with minimal investment—prefer to use fully-managed infrastructure, as it enables them to easily scale ML workloads without infrastructure management requirements. **We recommend managed infrastructure if you don't have dedicated infrastructure management and/or AI research teams.**

Some companies prefer to build their own custom ML platform to meet highly specialized ML requirements and frameworks, and to satisfy specific resource utilization and infrastructure strategies. In this instance, companies that modernize their architecture on containers use GKE for scale and reliability. GKE autoscaling automatically matches workload demands and can support 15,000 node clusters—the largest in the industry—for wide-scale distributed use cases.



Did you know?



A single company can choose both Vertex AI's managed platform and custom built options on GKE and GCE. In fact, most do. Our large customers have many teams with different operational modes and needs—one team may prefer the simplified, end-to-end experience provided by Vertex AI, while another may need more precise control over resources.

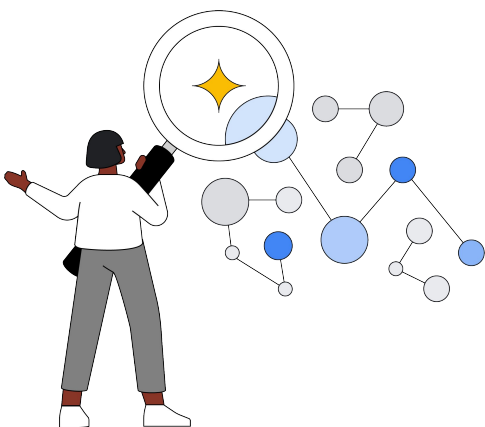


Learn how [Etsy uses Vertex AI](#) and custom built options.

Choosing your AI infrastructure

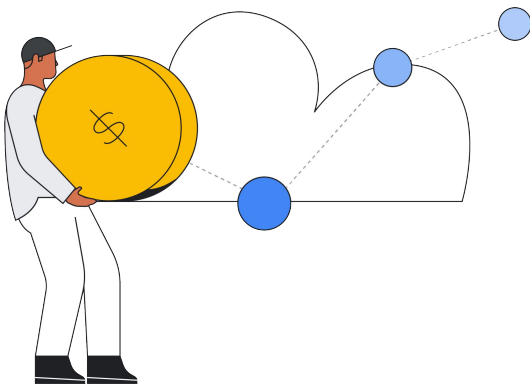
Service	Description	A good match if you:
Fully-managed AI infrastructure with Vertex AI	An end-to-end ML platform with fully-managed tools, workflows, and infrastructure.	<ul style="list-style-type: none"> Want minimal infrastructure management Want strong integration with Data Cloud (BigQuery, Spark, Dataflow) and built-in MLOps tooling Want to access state-of-the-art AI from Google Research (e.g. AutoML, NAS, TabNet, Feature Transformation Engine) Don't want to invest in an engineering and operations team to build and manage an in-house infrastructure platform
Custom build with OSS and GKE	<p>Ray/Kubeflow/Flyte/MLflow/Feast on GKE</p> <p>TPUs & GPUs on GCE/GKE</p>	<ul style="list-style-type: none"> Have a large in-house ML infrastructure team with strong Kubernetes and infrastructure knowledge Want to adopt fully OSS and portable ML infrastructure Have an AI research team looking for lowest level access to AI accelerators to optimize runtime and code

Once you have a customized solution in place, you're ready to start leveraging key features and benefits. Let's dig in.



DEEP DIVE

Optimize performance and costs with fully-managed AI infrastructure



With [76% of enterprises](#) saying that AI projects are their top priority, it's never been more important to re-architect or re-skill your teams to support AI workloads. With Vertex AI's fully-managed infrastructure, you can innovate faster and more easily—all while optimizing costs.

Vertex AI provides purpose-built AI infrastructure for **large-scale training** and **low-latency serving**, with access to a wide variety of compute, storage, and networking capabilities to meet your unique performance and budget needs. And because the infrastructure is fully managed, you don't need infrastructure expertise to get the most out of your AI workloads.

Let's dive deeper into what you can do with Vertex AI.

1. Accelerate large-scale training



Manage training jobs

Often, the development efficiency of data scientists and ML engineers is limited by how frequently they can train their deep learning models to incorporate new features, improve prediction accuracy, or adjust for data drift. Vertex AI's fully-managed training service eliminates the need to provision and manage clusters, enabling teams to **accelerate model development by 5x** and scale ML workloads more efficiently. Users without infrastructure expertise can submit and forget training jobs with a Python SDK, get out-of-the-box queue management, and access on-demand GPUs and TPUs with built-in hyperparameter optimization.



Optimize distributed training

As deep learning models become more complex and the size of training datasets grows, training time has become one of the key bottlenecks in the development and deployment of ML systems. To speed up training of large models, many engineering teams are adopting distributed training using scale-out clusters of ML accelerators. Yet limited network bandwidth between nodes makes it hard to optimize the performance of this training. Vertex AI Reduction Server provides a novel architecture that optimizes bandwidth and latency of multi-node distributed training on NVIDIA GPUs, with benchmarks for natural language models showing up to **40% reduction in training time and cost.**



Store and access large data sets

Data feeding into accelerators can often be a bottleneck. Vertex AI provides first-class support for Cloud Storage and NFS file store, making it easy to ingest data into demand training jobs as file systems. With built-in NFS and Cloud Storage FUSE support for custom training jobs, you gain an easy way to store and access large datasets, with less of the cumbersome work of moving training data around. You can also execute training jobs faster by eliminating the data download steps, and enable data streams over the network with higher throughput compared to alternative storage solutions.



2. Deliver low-latency serving



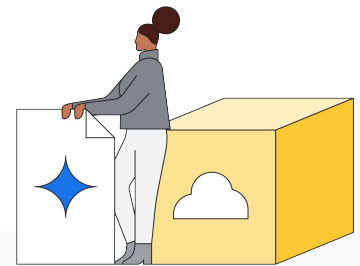
Optimize proprietary workloads for serving

From product recommendations, to fraud detection or route optimization, low-latency predictions are vital in many ML tasks. With Vertex AI, you gain access to a new runtime that uses the same technologies and model optimization techniques that are used internally at Google to optimize serving TensorFlow models on the Vertex AI Prediction service. They can be incorporated into your serving workflows without any changes to your training or model saving code, leading to faster predictions at a lower cost compared to the open source pre-built [TensorFlow serving containers](#). Benchmarks show improvements of **up to 8x for throughput and 6x for latency** for tabular data, and material improvements for BERT-based models.



Real-time inference

Real-time inference is ideal for inference workloads where you have real-time, interactive, low-latency requirements. You can deploy your model to Vertex AI and get an endpoint that can be used for inference. These endpoints are fully managed and support auto-scaling with a wide section of CPUs and GPUs. Dataflow is deeply integrated with Vertex AI for inference workloads, with support for both CPUs and GPUs.



Vodafone is supercharging AI and ML at scale

Vodafone's Commercial team has a strong history of using AI and ML to drive business results. Yet as demand grows, embedding AI and ML into the fabric of the organization gets harder—as does rapidly building and deploying ML use cases at scale in a highly regulated industry. To solve these issues, Vodafone built a scalable, unified ML platform called AI Booster on Google Cloud. Now, moving a model from proof of concept (PoC) to production **can now be as little as four weeks, which is 80% faster.**

Thanks to AI Booster, data scientists and ML engineers are driving greater value and collaborating on innovative solutions. And, supported by instructor-led and on-demand learning paths within Google Cloud, a new culture of experimentation and learning is being fostered at Vodafone.



[Read Vodafone's story](#)

3. Drive value with a strong MLOps governance layer

Deploying models is only half the battle for companies that want to deliver high-value MLOps at scale. Vertex AI's MLOps capabilities help companies automate, standardize, and accelerate ML processes, reducing the work needed to maintain model performance in production—which in turn helps reduce costs.



Automate pipelines

Automating training workflows helps data scientists orchestrate ML workflows in a serverless manner, with faster experimentation and model training. With Vertex AI pipelines, you can automate the end-to-end ML workflow from data ingestion to model deployment.



Monitor and explain models

Once models are deployed, they must be maintained, monitored, and re-trained—which is difficult to manage at scale. To help you maintain a model's performance, Vertex AI's model monitoring and explainability capabilities ensure models stay healthy and are re-trained or re-deployed as they erode.



Retrieve features with low latency

Building a feature store, or feature engineering, is critically important when building an ML model. The quality of your features directly influences your model predictability, and often needs a few iterations until a model reaches an ideal level of accuracy. Vertex AI provides a fully-managed feature store that lets you define features, use batch ingestions and streaming ingestion, and retrieve features with low latency.

Did you know?



Data scientists and developers can easily spend 60% of their time designing and creating features. With the Vertex AI feature store, teams can use that time to focus on innovating code.

Myth

If I use an end-to-end managed service, I'll be locked-in and won't be able to use other tools.

Reality

Vertex AI's managed services are modular so you can use them alongside your existing ML framework.

DEEP DIVE

Support diverse workloads with flexible and open infrastructure that meets your needs



Flexible building blocks

Train your models faster and more cost-effectively

Armed with the right **compute, storage, and networking building blocks**, your teams can iterate faster, train more models, and increase prediction accuracy—while reducing costs at the same time. With Google Cloud, you can tick all these boxes while knowing you're using the same foundation that Google uses for planet-scale applications like Photos, YouTube, and Gmail. That is, exabyte-scale and highly performant services built and refined over the last 20 years.

Let's take a look at these building blocks in a little more detail.

COMPUTE



Compute power for every application and workload

Increasingly, companies need purpose-built AI infrastructure that can scale up to take advantage of multiple accelerators within a single server and scale out to combine many servers distributed across a high-performance network. Google Cloud delivers, with purpose-built AI accelerators that meet the most demanding real-world workloads at scale, while meeting budget and time-to-solution requirements.

You can choose from GPUs, CPUs and TPUs (Google's custom-developed hardware accelerator, which we use to power our AI workloads), with all options purpose-built for AI.

01

CPUs

- Several cores
- Serial processing
- Quick prototyping on simple fast training models
- Small models that work with effective batch sizes

02

GPUs

- Thousands of cores
- High data throughput
- Massive parallel computing

03

TPUs

- Matrix based workloads
- Models with matrix computations
- Models with no custom TensorFlow operations
- Models that train for weeks or months



Optimize costs

Google Cloud's NVIDIA A2 VM instances with A100 GPUs are purpose-built to optimize AI and ML workloads. Companies can rightsize infrastructure to meet demand, which improves utilization across workloads and helps save money. You can also partition GPUs into multiple instances to optimize training for your largest and smallest workloads. Leveraging GPU time-sharing through GKE can improve utilization.



Deliver low latency with greater flexibility

NVIDIA T4 TensorCore GPUs are designed for low-latency inference with a larger global footprint, which keeps compute power closer to you.

T4s have the flexible ability to accelerate a wide variety of workloads including high performance computing, inference, ML, data analytics, and graphics. And T4 GPUs have more VM shape flexibility so you can customize the VM to deploy and pay for only what you need.



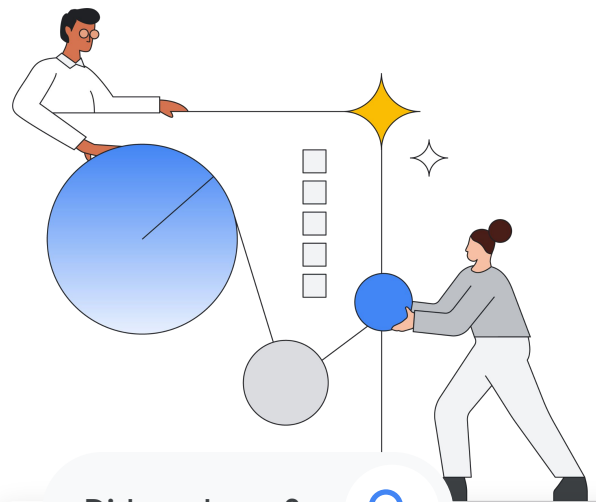
Easily handle large-scale workloads

Cloud TPUs are specifically designed to optimize performance per dollar at scale. Thanks to the high-speed Inter-Chip Interconnect (ICI) and XLA compiler optimizations such as compute communication overlap, Cloud TPU can deliver and sustain [performance efficiency](#) when you scale from one to hundreds of TPU chips—helping minimize time-to-accuracy for larger neural network models at low cost. Cloud TPUs support PyTorch, Jax, and TensorFlow, which all leverage XLA compiler to deliver exceptional performance and an easy developer experience thanks to TPU VM and TPU profiling tools.

Did you know?



Google Cloud is the only cloud provider that has up to 16x A100-40GB GPUs that can attach to a single VM for increased throughput.



Did you know?



Google Cloud's ML cluster with Cloud TPU v4 Pods is one of the fastest, most efficient, and most sustainable ML infrastructure hubs in the world.

STORAGE



Cost-effective, flexible storage solutions

With the help of ML and AI, organizations of all sizes and across all industries are using cloud storage to transform data into business value and provide enhanced customer experiences. As AI workloads grow, however, it can become harder to balance cost, capacity, performance, and business value over time.



Choose your storage solution

Depending on your source data format and application needs, we offer a variety of high performance, scalable, and cost-effective storage solutions—including Filestore for small files and I/O intensive application needs; Cloud storage for larger objects and sequential workloads; and Persistent Disk and I/O intensive block storage.



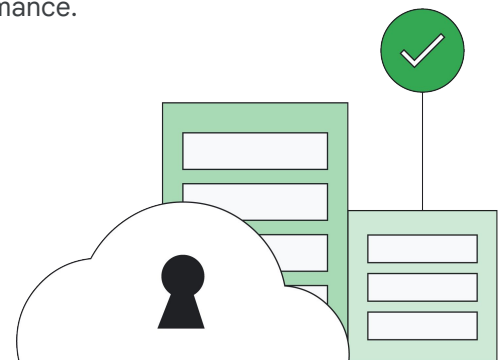
Choose your location

Availability and the physical location of data is an important consideration with storage. Cloud storage offers regional, multi-region, and dual region options—all with strong consistency of replicated data. We also offer an optional 15 minute recovery point objective (RPO) SLA with our dual region configuration.



Optimize costs

No matter which regional deployment you choose, you can optimize costs by enabling the Cloud Storage Autoclass feature which, by policy, dynamically migrates objects across different storage classes based on access frequency, with no impact on performance. Cloud storage provides low latency time to first byte, measured in milliseconds to seconds regardless of storage class, so you can optimize costs without compromising application performance.



AI IN ACTION

Google Cloud is helping CERN researchers **analyze powerful data** and **uncover the secrets of our universe**.



In 2012, the analysis of the full dataset took more than 24 hours. When we reran the full analysis on Google Cloud, it took just over five minutes. Through Google Cloud, we were able to access 25,000 cores very quickly, run the analysis, and then shut everything down again.”

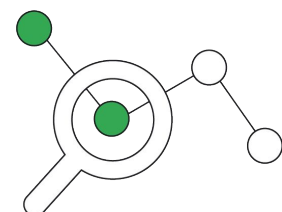
Ricardo Rocha
Computing Engineer, CERN



[Read Cern's story](#)

About CERN

The European Organization for Nuclear Research (CERN) uses the world's most complex scientific instruments, including the Large Hadron Collider, to study subatomic particles and advance the boundaries of human knowledge by delving into the smallest building blocks of nature. Founded in 1954, CERN was one of Europe's first joint ventures and now has 23 member states.



NETWORKING



Networking to help train your ML models faster

The larger and more complex the model is, the longer the training is going to take. Networking solutions help train ML models faster. With our state-of-the-art dedicated Google Cloud backbone network and [Jupiter data center fabric](#), our infrastructure is fundamentally different from other providers.



Consistent low-latency

Our network fabric, disaggregated compute and storage, optimized network virtualization stack (Andromeda), and compact placement policies deliver better latency and throughput for your ML applications. Optical circuit switching (OCS), wave division multiplexing (WDM), and software-defined networking (SDN) combine to form our ultra-low-latency Jupiter network fabric.



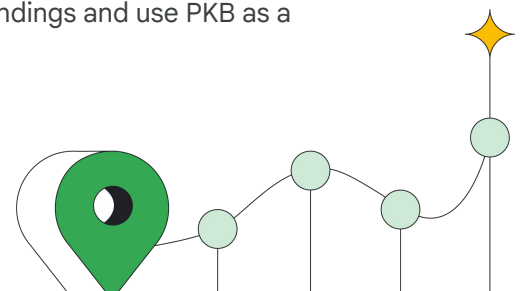
High performance

Our network fabric supports 6 Pb/sec of datacenter bandwidth, a six-fold increase from 2015. With our [Tier 1 networking](#), you can increase the per-VM egress throughput limits to up to 100 Gbps depending on the chosen VM type. Our [latest C3 VMs](#), announced in preview in October 2022, go even further by providing up to 200 Gbps low-latency networking and line rate PSP encryption of traffic. Such high performance networking is critical for large, distributed ML workloads with heavy internode communications.



Transparent validation

Every customer can validate Google Cloud's performance both for their own projects and as a global network. [Network Intelligence Center Performance Dashboards](#) provide network latencies and packet loss metrics for both performance monitoring and deployment planning. Our cloud benchmark efforts as a public and open source project, [PerfKit Benchmark](#) (PKB), so that others can replicate our findings and use PKB as a standard for cloud performance evaluation.





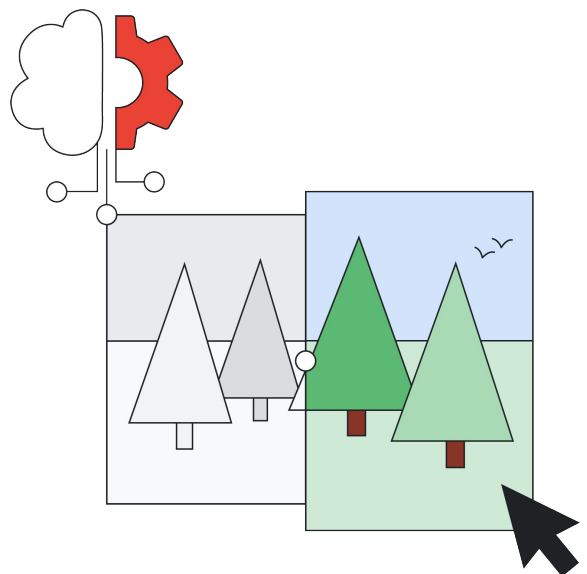
Let's enhance uses GKE to power AI-based photo editing

Let's Enhance is a computer vision company that improves the quality of user-generated photos using AI-based features. Yet building and deploying this AI-enabled service at global scale presented a huge technical challenge spanning model building, training, inference serving, and resource scaling. They needed an infrastructure that would be easy to manage and monitor, could deliver real-time performance to end customers wherever they are, and scale as user demand peaks—all while optimizing costs.

To support their growing user base, Let's Enhance chose to deploy their AI-powered platform to production on GKE, with Google Cloud providing the compute, infrastructure management, and inference management needed for rapid growth and scale. Since implementation, Let's Enhance observed a **40% average throughput improvement and 34% cost savings using MIG-enabled A100 GPUs**.



[Read the Let's Enhance story](#)



An open ecosystem

Make the most of ML with open source software

At Google, we believe anyone should be able to quickly and easily turn their AI idea into reality. OSS is increasingly important here, with a huge impact on the pace of innovation in AI and ML ecosystems.

Over the last two decades, ML has transformed Google services like Search, YouTube, Assistant, and Maps—and the driving force behind these transformations is our ‘open first’ approach through investments in projects and ecosystems like TensorFlow, Jax, and PyTorch. Today, our AI infrastructure services use the most popular ML frameworks—so you can keep using whichever framework you prefer, without constraint.

Did you know?



ML frameworks make it much easier to build ML models by allowing developers to focus on the overall logic and dataflow of their model.

Google proudly supports the future of open source AI tools

A

Thanks to our OSS contributions to TensorFlow, JAX, TFX, MLIR, KubeFlow, and Kubernetes—as well as sponsorship for critical OSS data science initiatives like Project Jupyter and NumFOCUS— Google is the leading Cloud Native Computing Foundation (CNCF) contributor.

B

Google collaborates with AI/ML industry leaders on the OpenXLA Project, an ecosystem of ML compiler and infrastructure technologies that let ML developers build their models on leading frameworks and execute them with high performance across hardware backends (GPU, CPU, and ML accelerators).

C

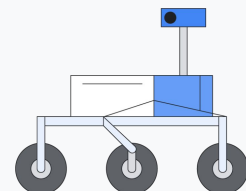
Google invests heavily in third-party platforms to make it easier for customers to use the ML platform of their choice. And, as a founding board member of the newly formed PyTorch Foundation, we will deepen our open source investment to deliver on the Foundation’s mission to drive adoption of AI tooling by building an ecosystem of open source projects with PyTorch.



Did you know?

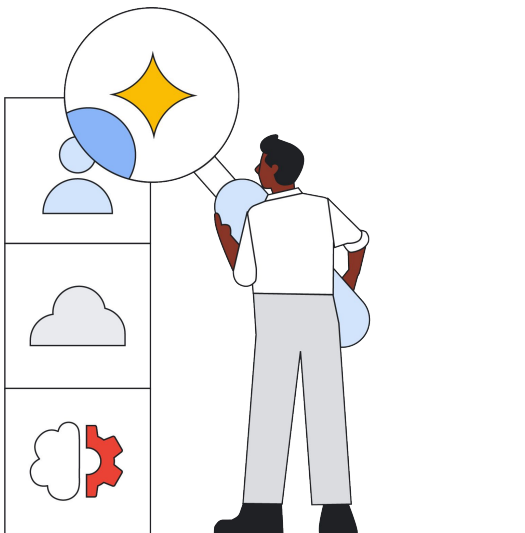


The **model that powered Mars rover Curiosity was deployed on TensorFlow**, Google’s open source ML framework. Part of NASA’s Mars Science Laboratory mission, Curiosity is the largest and most capable rover ever sent to Mars.



DEEP DIVE

Innovate faster with state-of-the-art AI solutions



To help every business drive true value through AI, Google is committed to providing an easy path to adopting leading-edge solutions. As with any cutting-edge technology, supporting infrastructure can often lag behind and hamper adoption. With Google Cloud's state-of-the-art AI solutions, teams can run the newest innovations from Google Research and partners on familiar platforms using Vertex AI.

Google's research delivers features which are scaled up from **discovery to enterprise-readiness** within two years. Vertex AI makes it far easier for data scientists to run these new innovations with optimized infrastructure and software paired together, so they can build, train and deploy more easily.

Recent innovations from Google Cloud and partners are helping to solve some of the world's biggest challenges.



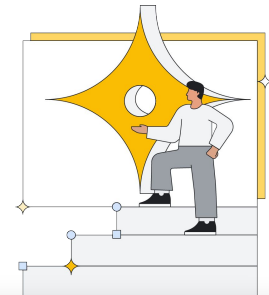
Ensuring accurate AI at scale

Data scientists choose models based on various tradeoffs when solving ML problems that involve tabular (structured) data. Decision trees are popular here, because they are easy to interpret, fast to train, and can obtain high accuracy quickly from small-scale datasets. On the other hand, deep neural networks offer superior accuracy on larger datasets, as well as the benefits of end-to-end learning—yet they are black-box and difficult to interpret.

[TabNet](#), an interpretable deep learning architecture developed by Google AI, combines the best of both worlds: it is explainable, like simpler tree-based models; and can achieve the high accuracy of complex black-box models and ensembles. TabNet on Vertex AI provides fully-managed, optimized, and scalable pipelines geared for efficient scaling to massive tabular datasets. Yielding better accuracy for real-world data challenges, it's ideal for a wide range of tabular data tasks where model explainability is just as important as accuracy—such as financial asset price prediction, fraud/cyber attack/crime detection, retail demand forecasting, user modeling, credit/risk scoring, diagnosis from healthcare records, and product recommendations.



[Read TabNet's story](#)



Transforming customer experiences

Recommender systems are the lifeline of retail businesses. With rapid growth in the scale of retail datasets, deep learning (DL) recommender models deliver advantages over traditional methods by capitalizing on large amounts of training data. However, developing and operationalizing large-scale DL recommender models can be very challenging. Recommender system framework NVIDIA Merlin and Vertex AI help streamline and accelerate the data pre-processing, training, and serving of large-scale DL recommenders.

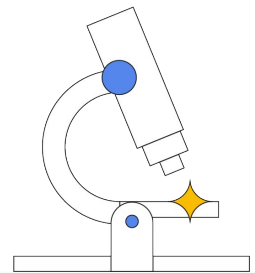
NVIDIA Merlin and Vertex AI overcome the technical challenges of delivering large-scale SOTA recommender systems that meet customers' expectation of high speed and accuracy. They can deliver a 100x improvement in training step time, and can implement a full two-tower model with 100x fewer lines of code.



Improving the health of humanity

As researchers seek cures for diseases and solutions for the big problems facing humankind—like antibiotic resistance, microplastic pollution, and climate change—they will benefit from fresh insights into the structure of proteins. Proteins are like tiny exquisite biological machines. The same way that the structure of a machine tells you what it does, so the structure of a protein helps us understand its function.

DeepMind, an AI research organization within Alphabet, is sharing a trove of information that doubles humanity's understanding of the human proteome, and reveals the protein structures found in 20 other biologically-significant organisms, from E.coli to yeast, and from the fruit fly to the mouse. The AlphaFold system advances this area of research by helping data scientists and other researchers to accurately predict protein geometries at scale. Batch inference with Vertex AI lets you efficiently run AlphaFold inference at scale with several optimizations by parallelizing independent steps. The solution automatically provisions and deprovisions the compute resources required for each step and runs on the most optimal hardware platform. And it provides a robust and flexible experiment tracking approach that simplifies the process of running and analyzing hundred of inference workloads.



DeepMind enhances Google Cloud's core AI solutions

Over the past few years, DeepMind has partnered with Google Cloud to apply its AI research to the core solutions used by our customers. This positive partnership has brought a number of innovative projects to life—including optimizing document understanding, enhancing the value of wind energy, and making it easier than ever for scientists to use AlphaFold.



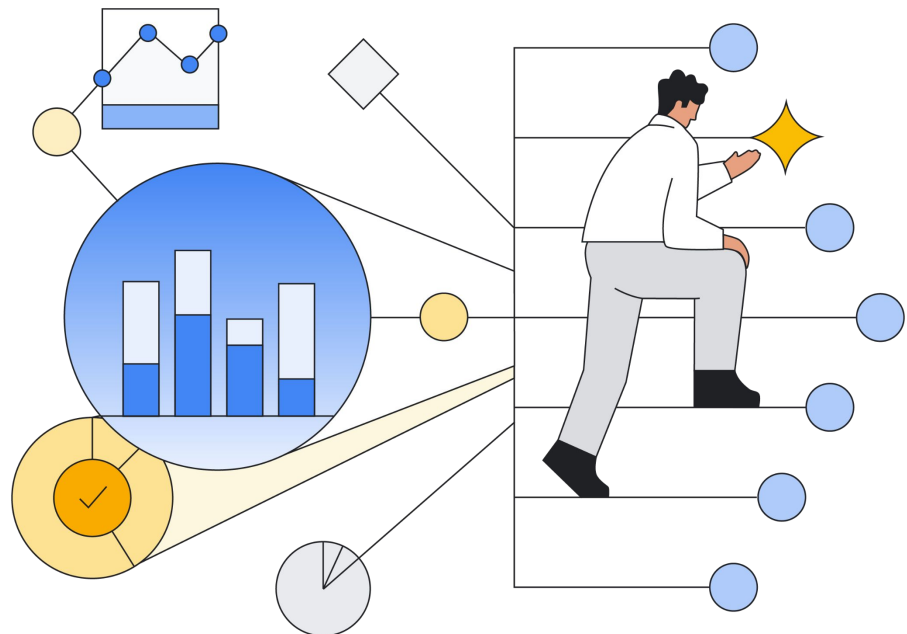
[Learn more about the partnership](#)



Unlocking natural language capabilities

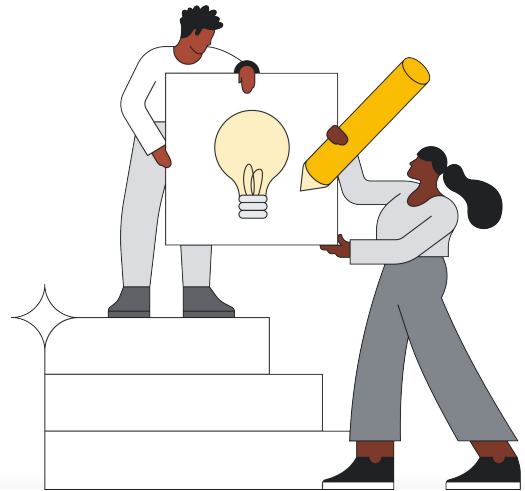
Innovations in natural language across machine translation, question answering, text summarization, and more continue to unlock new use cases for enterprises in all industries. Google Research publishes research papers, tools, and open source code to make building and consuming natural language models more accessible to developers and data scientists. The [T5 text-to-text transformer approach](#), published by Google Research in 2020, provides a unified and flexible framework capable of solving a large variety of natural language tasks. Google Cloud users can leverage the T5 and [T5X frameworks on Vertex AI](#) to take advantage of streamlined experimentation, development, and deployment of natural language processing solutions at scale on Google Cloud's TPUs.

T5X is built as a JAX-based library for training, evaluating, and inferring with sequence models. T5X is streamlined, modular, and composable. You can implement pretraining, fine-tuning, evaluating, and inferring by configuring reusable components deployed on Vertex AI.



Are you ready to drive new AI innovations on Google Cloud?

Your journey to building high performance, cost-effective AI infrastructure on Google Cloud starts here:



[Contact sales about AI infrastructure](#)



Get started on your own

- [Build with Vertex AI's fully-managed AI infrastructure](#)
- Build your own ML platform with
 - [Google Kubernetes Engine \(GKE\)](#)
 - [Deep Learning VMs](#)



Explore Coursera lab courses and code labs

- Blog: [Take your ML models from prototype to production](#)
 - Code lab: [Training custom models on Vertex AI](#)
 - Codelab: [Getting predictions from custom trained models](#)
 - Codelab: [Distributed training on Vertex AI](#)

Turn your cloud vision into a reality with this free migration and modernization assessment [→](#)