Google Cloud

jornadas FCCN

Google Cloud for Researchers
Jornadas FCCN

Luís João
Head of Public Sector
Google Cloud Portugal

April 17th, 2024

Google Cloud Next

All 218 things we announced at Google Cloud Next ‘24 – a recap

April 13, 2024

Gemini for Google Cloud

Software Development
Accelerate software delivery

Gemini Code Assist

Application Lifecycle
Efficiently manage cloud applications

Gemini Cloud Assist

Security
Elevate security expertise

Gemini in Security

Data Analytics
Fast-track data analysis

Gemini in BigQuery

Business Intelligence
Automate data insights

Gemini in Looker

Databases
Supercharge database development & management

Gemini in Databases

90%

of generative AI unicorns are Google Cloud customers

Customer Segmentation

Foundation Model Producers

- Training Speed/Cost
- Choice of AI Infrastructure
- Leading AI Frameworks

Foundation Model Tuners

- Tuning Speed/Cost
- Choice of AI Infrastructure
- Data & Storage Integration

Foundation Model Consumers

- Low-latency Serving
- Large-scale Inference
- Cost per Inference

The most comprehensive AI Infrastructure

Architected for Scale, Speed, Efficiency and Availability

Accessible to all Developers

- Foundation Model Consumers
- Foundation Model Tuners
- Foundation Model Producers

Vertex AI

- Search
- Conversation
- AI Solution: CCAI
- AI Solutions: Risk, etc
- Model Garden
- AI Platform
- ML Ops
- Foundation Models

AI Hypercomputer: *Optimized IAAS*

Flexible Consumption

- Dynamic Workload Scheduler
- On Demand
- CUD
- Spot

Open Software

- JAX, TensorFlow, PyTorch
- Multislice Training, Multihost Inference, XLA
- Google Kubernetes Engine & Compute Engine

Performance-Optimized Hardware

- Compute (GPUs, TPUs)
- Storage (Block, File, Object)
- Networking (OCS, Jupiter)

Ecosystem Partners

AI2, cohere, ANTHROP\C, Midjourney, stability.ai, osmo, Lightning AI, Hugging Face, anyscale, Weights & Biases, Snorkel, slurm

A leader in AI infrastructure

Google receives 5 out of 5 in 17 of 19 evaluation criteria

- Solution Architecture
- Solution Ecosystem
- Data Workloads
- Training Workloads
- Inferencing Workloads
- Management Tools
- Developer Tools
- Fault Tolerance
- Efficiency
- Vision
- Roadmap
- Innovation
- Partner ecosystem
- Pricing Transparency
- Supporting services and offerings
- Number of customers
- Revenue

FORRESTER®

The Forrester Wave™:
AI Infrastructure Solutions, Q1 2024

*A gray bubble or open dot indicates a nonparticipating vendor.

Source: Forrester Research, Inc. Unauthorized reproduction, citation, or distribution prohibited.

Boosting ROI for Large Scale ML projects

Delivering ML Productivity Goodput with end-to-end workflow optimizations

ML Productivity Goodput = Scheduling Goodput x Runtime Goodput x Program Goodput

*How often does an application have **all necessary resources** to make progress?*

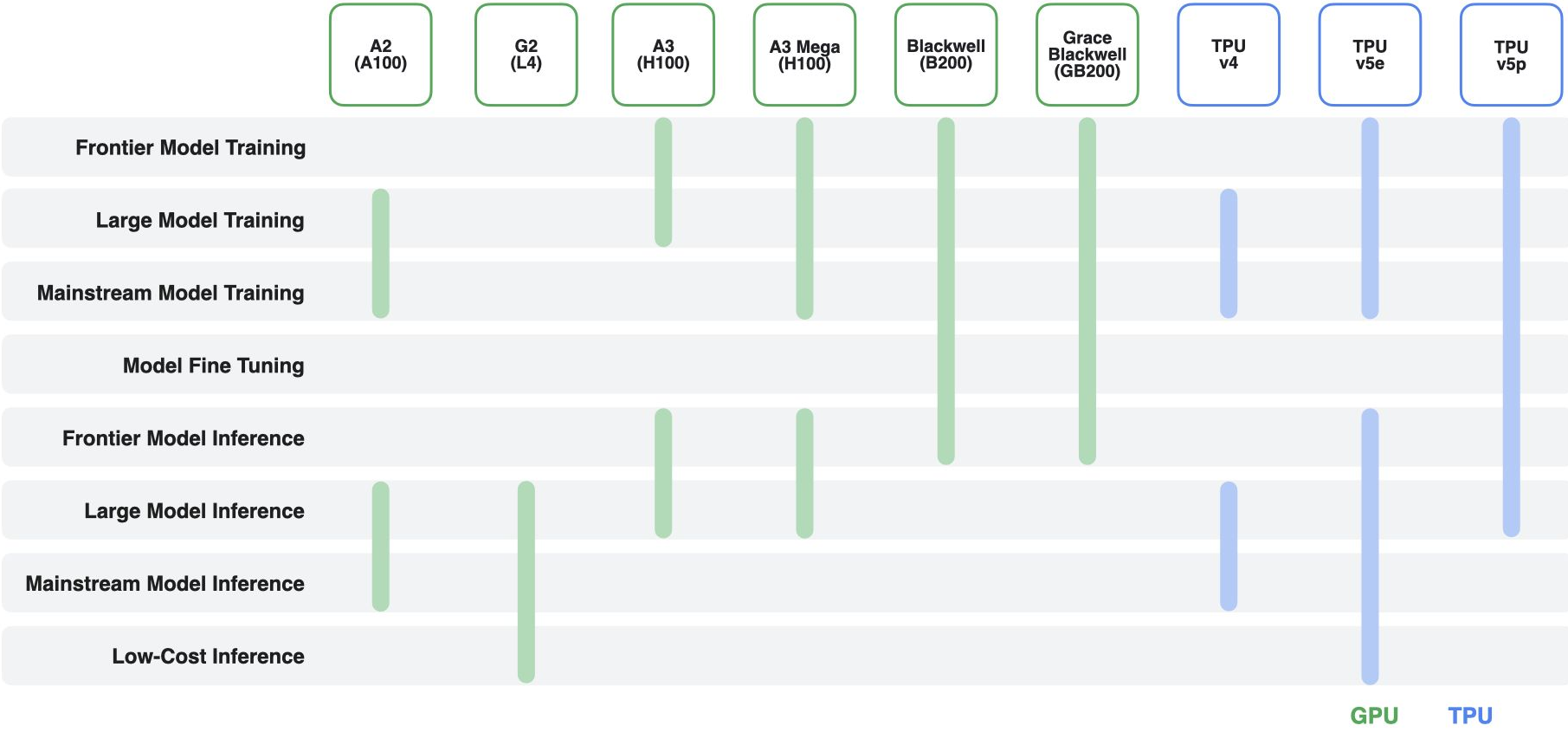
*Of the time that an application has all necessary resources, how often is it **making progress?***

*Of the time that an application is making progress, how often is it **perfectly efficient i.e., close to roofline**?*

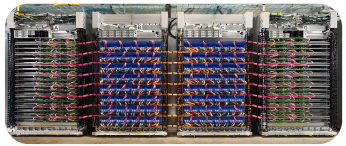
Example features:	Cloud Scheduler	ML Runtime & Auto-Checkpoint	ML Compiler

When designing our state-of-the-art ML infra, we are focusing on maximizing ML Productivity Goodput per $ by optimizing Scheduling, Runtime and Program software

GPUs and TPUs for every AI use case



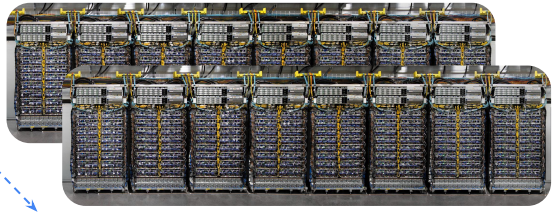
Rapid Innovation with Cloud TPUs



Cloud TPU v2

- Domain-specific AI supercomputing
- 256 chips distributed shared memory

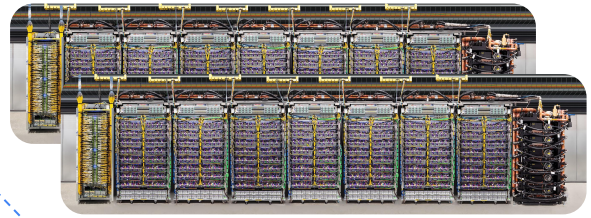
8x



Cloud TPU v4

- Optically reconfigurable 3D Torus
- 4k chips with distributed shared memory

20x



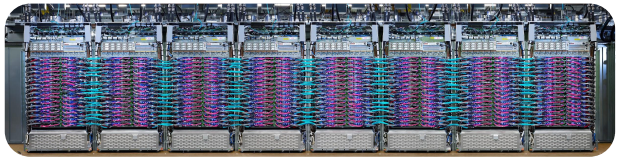
Cloud TPU v5p

- Programmable Sparsecores for embeddings
- 9k chips with distributed shared memory

2018 — 2020 — 2022 — 2023 — 2024

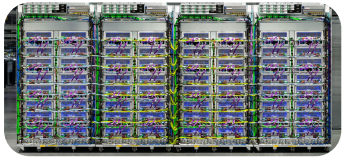
Cloud TPU v3

- Liquid cooling
- 1k chips distributed shared memory



Cloud TPU v5e

- Efficient and scalable training and serving
- 256 chips, horizontally scalable to 10s of k



Cloud TPU v5p

Powerful

Compress
SoTA AI model development
From Months to Weeks

2.8X Faster LLM Training
(vs TPU v4)

1.9X Faster Embeddings Training
(vs TPU v4)

Scalable

Effortlessly deploy
10s of K chips on the
Most scalable TPU to date

4x More Scalable
(vs TPU v4 pod)

World's Highest Bandwidth ICI
(4,800 Gbps/chip, 9k chip ICI domain)

Flexible

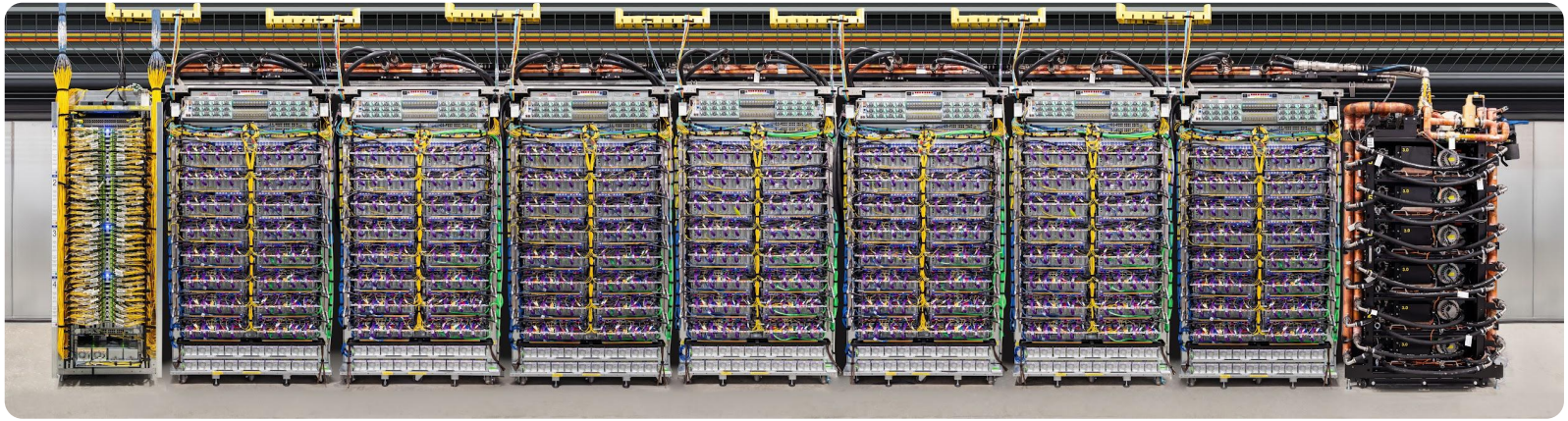
Accelerate the
full range of
AI model types & sizes

Full Range of AI Models
(LLMs, Recsys, Multimodal, MoE)

3X More HBM
(vs TPU v4)

Cloud TPU v5p

Powerful, Scalable, Easy-to-Use



Chip Specs

- Peak TFlops bf16: 459
- Peak TOPs Int8: 918
- HBM: 95GB @2,765 GBps
- Per chip ICI Bandwidth: 4800 Gbps

Pod Specs & Scaling

- 8,960 chips, Scalable to tens of Ks with multislice training
- 4.1 Exaflops (BF16) and 8.2 ExaOps (Int8)
- All Reduce Bandwidth per Pod: 5.3 PBps
- Bisection Bandwidth per Pod: 51 TBps

Availability

North America (us-east5-a)

Price

Starting at $1.89 / Chip-Hour

(with 3 year reservation)

High-performance inference on Cloud TPUs

Exported Model

TPU Compatible Model

TPU-Optimized Model

AI Frameworks

PyTorch

TensorFlow

JAX

Inference Converter

TPU conversion

Quantization

I/O shape optimization

Graph modification for GSPMD partitioning

XLA Compiler

OpenXLA

High-level fusion

GSPMD sharding

Low-level scheduling

Final hardware-specific optimizations & compilation

Model Serving

TensorFlow Serving

TorchServe

SAX

Voice of our customers

"Google Kubernetes Engine (GKE) allows us to run and optimize our GPU and TPU infrastructure at scale, while Vertex AI will enable us to distribute our models to customers via the Vertex AI Model Garden. Google's **next-generation AI infrastructure powered by A3 and TPU v5e will bring price-performance benefits for our workloads** as we continue to build the next wave of AI."

Tom Brown
Co-Founder

ANTHROP\C

"It's exciting to see the innovation of next generation accelerators in the overall AI landscape, including Google Cloud TPU v5e and A3 VMs with H100 GPUs. We expect both of these platforms to **offer more than 2X more cost-efficient performance** than their respective previous generations."

Noam Shazeer
CEO

character.ai

"Cloud TPUs have allowed Craiyon to **train models much faster and more efficiently,** which has led to a significant improvement in the quality of our AI-generated content. For example, we were able to gain the same performance on Cloud TPU v5e using only half the cores as that of the Cloud TPU v4 generation."

Boris Dayma
Founder

craiyon

"We're a huge fan of Google Cloud TPUs. Our speed benchmarks are demonstrating a 5X increase in the speed of AI models when training and running on Google Cloud TPU v5e. We are also seeing a tremendous improvement in the scale of our inference metrics, we can now process **1000 seconds with one real-time second for in-house speech-to-text and emotion prediction models - a 6x improvement.**"

Wonkyum Lee
Head of Machine Learning

Gridspace

"Google Deepmind and Google Research have had several successful training runs each using **many thousands of TPU v5e chips** including models for LLM use cases with **excellent scaling efficiency** - similar to TPU v4 generation - using Multislice scaling software"

Jeff Dean
Chief Scientist

Google
Google Cloud

Industry-leading Cloud GPU Platform

Optimized for your workload, powered by NVIDIA

Speed to Market

1st Cloud to launch P4 & T4 for accelerated ML inference

1st Cloud to launch A100 for highest perf ML training

1st Cloud to launch L4, purpose-built for Gen AI inference

Breadth of Offerings

G2 VM with L4 GPU to accelerate gen AI inference

A3 VM with H100 GPU for largest scale AI training

Built on Jupiter network, 40% less power, adaptive & flexible

Powerful & Productive

Leading contributor of OSS dev tools such as JAX & OpenXLA

GKE Enterprise improves team productivity by 45%

Train models 5x faster with Vertex AI

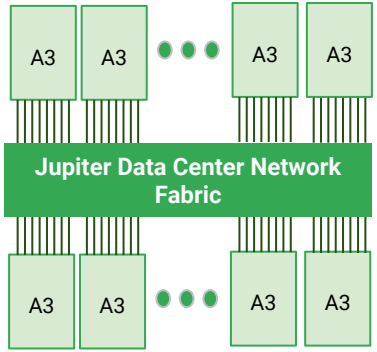
A3 VMs

10x network bandwidth and 3x training performance increase over A2

NVIDIA H100 GPU

26 exaFlops of AI Performance

Scale to 10s of Thousands of GPUs



Built on Google Jupiter Data Center Fabric

Use Cases

Large Model Training

Large Model Serving

Scientific computing

Trusted by Gen AI Unicorns

ANTHROP\C

character.ai

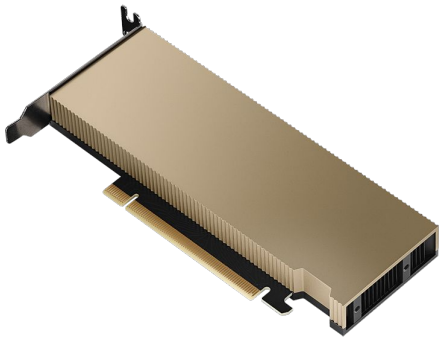
Midjourney

BENDING SPOONS

G2 VMs: Optimized for Gen AI and Graphics

2-4x performance boost over T4 GPU

NVIDIA L4 GPU



1st Cloud to launch NVIDIA L4 GPU

Use Cases

2.5X Gen AI Inference

3.3X Gaming

4X Digital Twin

Trusted by AI Partners

APPLOVIN

BENDING SPOONS

Snap

Volkswagen

Workspot.

Broad range of consumption models

Tailored to fit any customer workload

On-demand

- Best capacity availability
- Highest flexibility for bursting
- No preemption

Committed Use Discounts

Discount: **between 30-55%**

- Ideal for **steady state** workload
- Discount apply to **aggregated resource** within a region and machine family

Spot VMs

Dynamic discounts

- Lowest price
- Disruption tolerant workloads
- Graceful termination

(Shared) Reservations

- **Leverages same discounts** as applied to on-demand and CUD usage
- Ideal for **mission-critical workloads**, **business-critical events**
- Need to specify **count, machine-type and zones;**

Dynamic Workload Scheduler (DWS)

New obtainability capabilities for accelerators

Works across GCP

- Managed Instance Groups on GCE
- Batch on GCE
- GKE
- Vertex AI

Calendar Mode:
Job start times assurance with Future Reservations

Use Cases:
(re)training, recurring fine-tuning

GPUs

Flex Start Mode:
Optimized economics and higher obtainability for on-demand resources

Use Cases:
time flexible experiments, fine tuning, batch inference

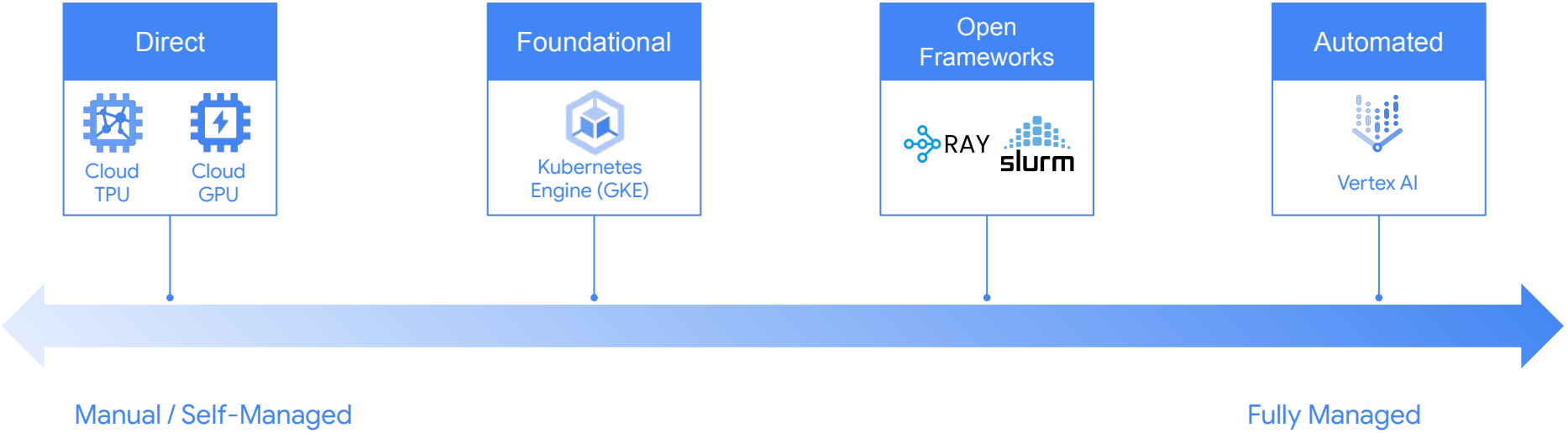
GPUs & TPUs

"The new DWS scheduling capabilities have been a game-changer in procuring sufficient GPU capacity for our training runs. We didn't have to worry about wasting money on idle GPUs while refreshing the page hoping for sufficient compute resources to become available."

- Sahil Chopra, Co-Founder & CEO, Linum AI

Workload Orchestration - Open, Flexible, and Accessible

Leverage TPU and GPU supercomputers your preferred way



Kubernetes Orchestration Built for AI Workloads

Google Kubernetes Engine

Google Console | GCloud CLI | GKE API | GKE Dashboard | Terraform

GKE Orchestration for TPUs & GPUs
Abstracts Accelerators, Networking, and Storage.
Enables efficient sharing of resources at scale.

- **Pre-flight Checks:** Ensure jobs run on validated and healthy TPU nodes.
- **Autoscaling**: Scale-up up or down to meet changing demands and drive savings.
- **Workload orchestration**: Grant guaranteed capacity for burst workloads via fair sharing, queuing, pre-emption, and prioritization.
- **Consistent Ops Environment**: A single platform for all AI/ML and other workloads.
- **Cost-optimized Profiles**: Configure GKE to save you money out-of-the-box.
- **Automatic Upgrades**: Minimize manual effort because GKE stays up-to-date.
- **Load balancing**: Easily distribute workloads to minimize customer latency.

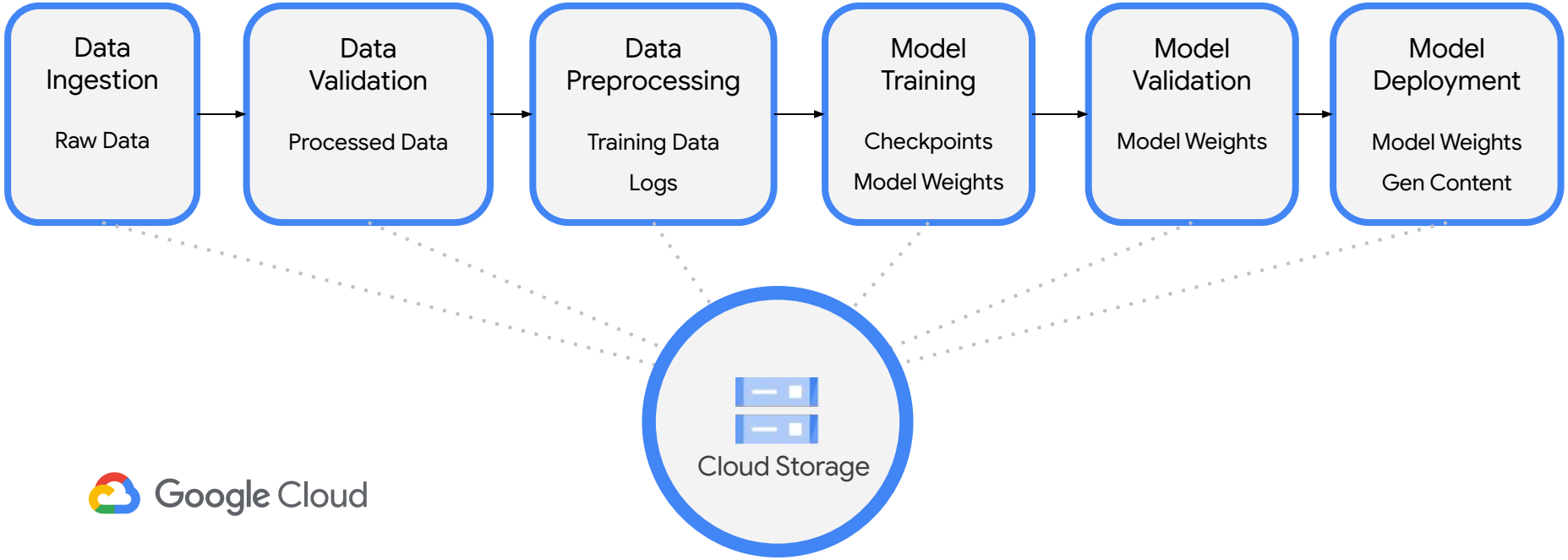
Cloud Storage is the source of truth for AI/ML data pipelines

Performant
- Petabytes of storage
- Tbps of throughput
- Billions of objects
- ms of latency

Reliable
- Redundant across zones/regions
- Industry only RPO SLA of 15 min

Secure
- Encrypted by default
- Integrated with IAM and KMS
- Granular permissions

Intelligent
- Automated storage selection
- Metadata insights reports
- Protection against deletes



Google Cloud

Cloud Storage is integrated into your overall Data Ecosystems

Build Intelligence for Complex Problems

Google Kubernetes Engine
Computes Engine
Vertex AI
GPU
TPU
Tensorflow
JAX
Pythorch

Process, analyze, predict

Managed services to transform data into insight

Cloud Dataflow
Cloud Dataproc
Google BigQuery

Data Management

Data governance at scale

Dataplex
Datacatalog

Cloud Storage

Data ingestion

Quickly and securely move your data into Cloud Storage

Cloud Pub/Sub
Transfer appliance
Dedicated interconnect
Partner interconnect

Data Access and Security

Low friction and secure access to data

Cloud IAM
KMS
DLP
Apigee API platform

Google Cloud

LangChain is supported across all Google databases

Rapid development | Interoperability | Enterprise-grade features

Now supported in:

- Cloud SQL for MySQL
- Cloud SQL for PostgreSQL
- Cloud SQL for SQL Server*
- AlloyDB
- Spanner
- Bigtable*
- Memorystore for Redis
- Firestore

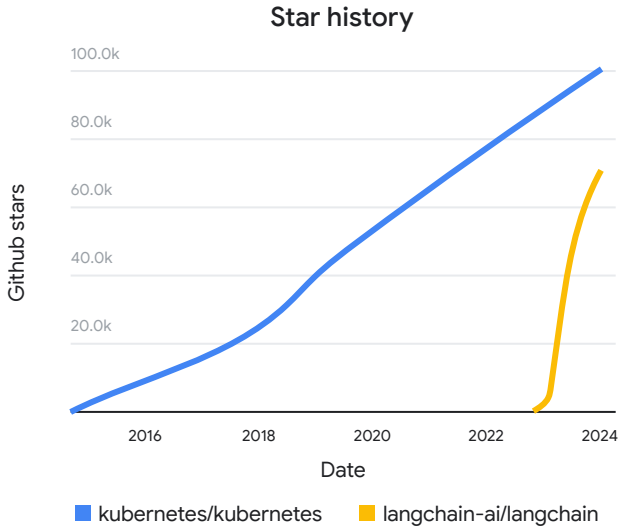
LangChain is the most popular OSS gen AI framework

GitHub repo

+700 different integrations

+2k contributors

~70k stars



* no vector store

star-history.com

Helping researchers at CERN to analyze powerful data and uncover the secrets of our universe

CERN analyzes petabytes of data per year, including from experiments on the world's largest particle accelerator. A joint project has shown how it's possible to burst this infrastructure with Google Cloud.

Google Cloud results

- Sped up terabyte-size workloads by reading data at 200 GB per second with Cloud Storage
- Compute power was scaled automatically, as needed, with Google Kubernetes Engine
- Used the public cloud for the public good by making more data open source for researchers, scientists, and educators

Researchers analyze 70 TB Higgs boson data in minutes

Tell us your challenge. We're here to help.

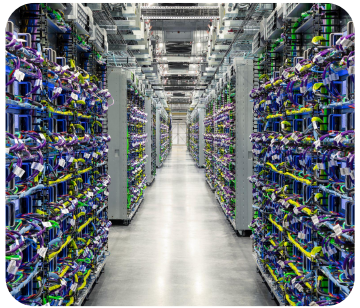
Contact us

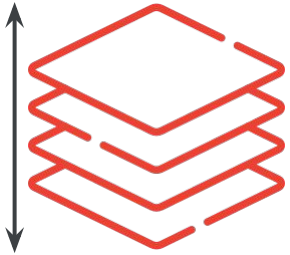
CERN openlab

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.

Why customers prefer Google Cloud for AI





Choose

Ultra performant AI supercomputers for any workload

TPU

&

GPU

Build

On Open & Comprehensive AI stack fueling GenAI revolution

Transformer

JAX

XLA

Deploy

Largest AI workloads with high reliability, availability, security, and goodput

Goodput

For ML Workloads

Secure

6 layers

Scale

On our massive high-density data centers with liquid cooling, Jupiter networking

37

Regions

200+

Countries

Sustainable

Reduce carbon emissions with the greenest cloud

100%

Renewable energy matched globally

1.10 PUE

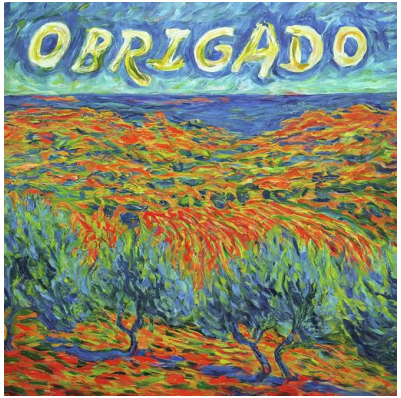
Gemini Advanced

Generate an artistic image with "Obrigado"

Sure, here is an artistic image with "Obrigado":

By Pablo Picasso

By Van Gogh



By Claude Monet

By Leonardo Da Vinci

Obrigado!

luisjoao@google.com