Google Innovation in HPC

FCCN Event

Google Cloud and High Performance Computing

Google innovative approach to HPC

HPC workloads aggregate computing power to solve large problems in science, engineering, or business – Generally achieved by using clusters of computers that are processing huge amounts of data

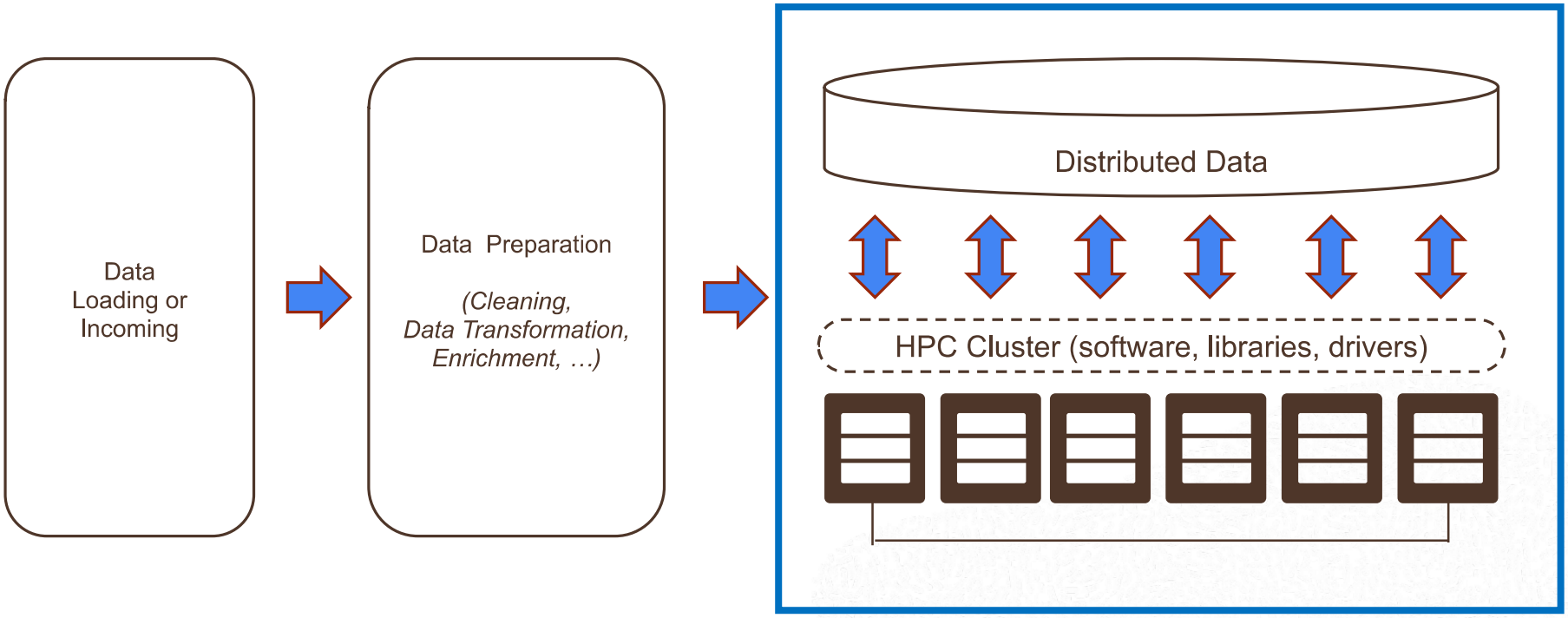
As usual Google has an innovative approach

- *BigQuery* as a highly distributed data management environment (GRID)
- GKE is the most scalable Kubernetes cluster (thousands of nodes)
- GKE uses NVidia technologies (e.g. P100, V100, A100 etc.)
- GCP provides data management technologies (e.g. Cloud Dataflow, Cloud Dataproc, etc.)
- Google has developed a high performance technologies to address Deep Learning workloads

Google Cloud

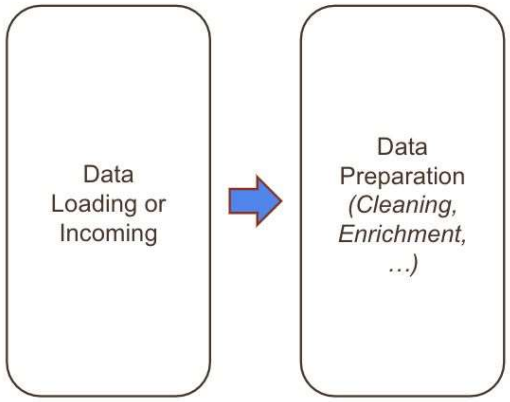
High Performance Computing Architecture

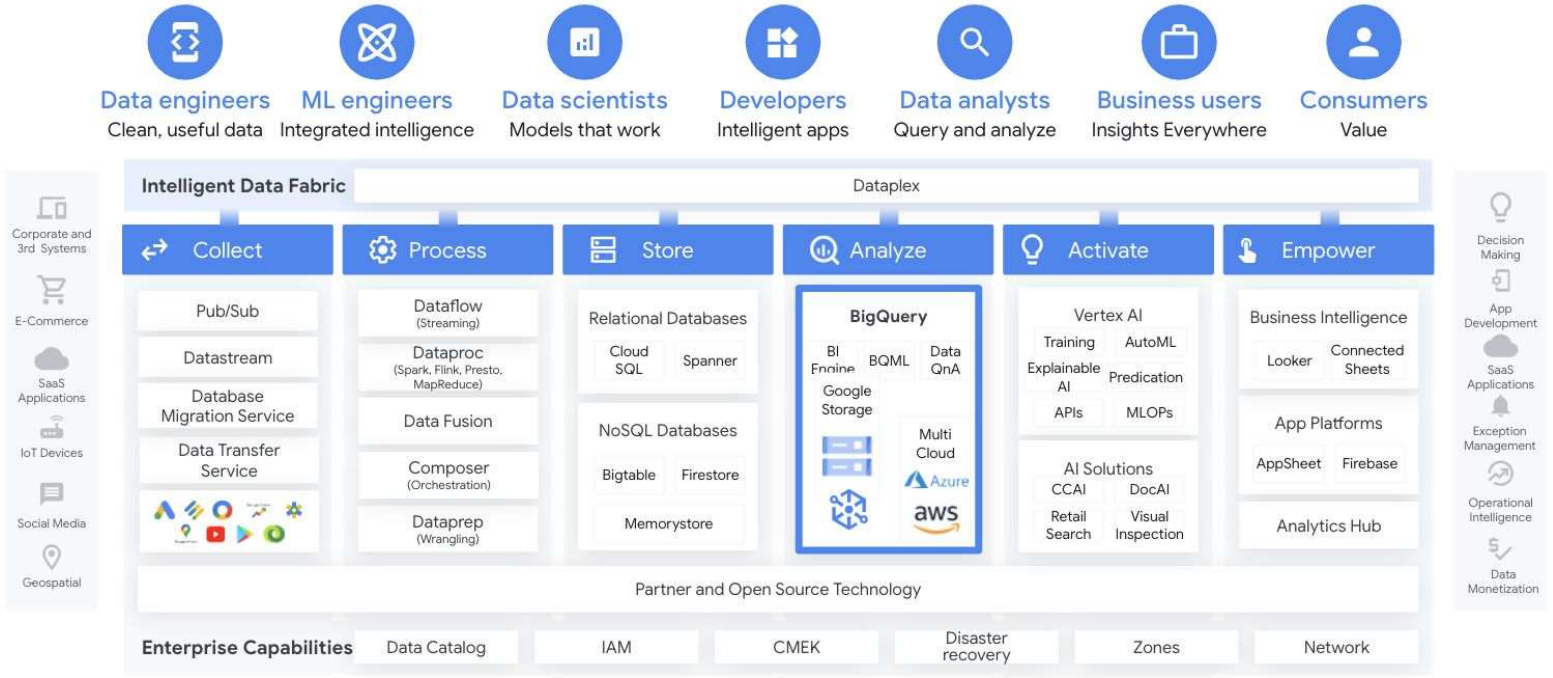
High Level HPC Data Lifecycle



High Performance Computing

Requires Data Preparation advanced tools

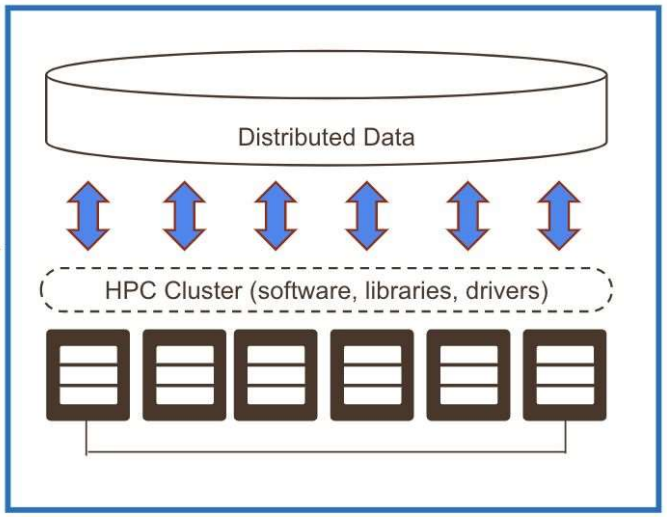




Google Cloud

High Performance Computing

High Performance Clusters



Highly Configurable Resources

Latest High Performance Processors

Customizable Instances

- 1 to 160 cores
- Up to 3844 GB RAM

Preemptible VMs

GPUs & TPUs

Managed Compute Paradigms

Pipelines API

Managed Instance Groups

Deployment Manager

Cloud Composer (Airflow)

Supports Various Workloads

MPI

Batch / HTC

Real Time

Mapreduce

Machine Learning

Works With Your Favorite Scheduler

Slurm

HTCondor

LSF

Grid Engine

Windows HPC Pack

More coming soon...

Google Cloud

The Value provided by HPC Cloud Solutions

- Really flexible to address Peak Demand for HPC resources
- Better technology refresh (on premise generally every 3-5yrs)
- Google Cloud provides best of breed data management solutions
- It's possible to use the right technology for each workloads (GPU, CPU, TPU, etc.)
- Hybrid Solutions (on premise and cloud HPC)

Google Cloud

Innovative HPC for AI - Cloud TPU

Custom ASIC by Google to train and execute deep neural networks

- Built for AI on Google Cloud
- Fast, iterative development
- Offers proven, Google-qualified reference models, optimized for performance, accuracy, and quality

At Google, the state-of-the-art capabilities you see in our products such as Search and YouTube are made possible by Tensor Processing Units (TPUs), our custom machine learning (ML) accelerators.

At 9 exaflops of peak aggregate performance, we believe our cluster of Cloud TPU v4 Pods is the world's largest publicly available ML hub in terms of cumulative computing power, while operating at 90% carbon-free energy

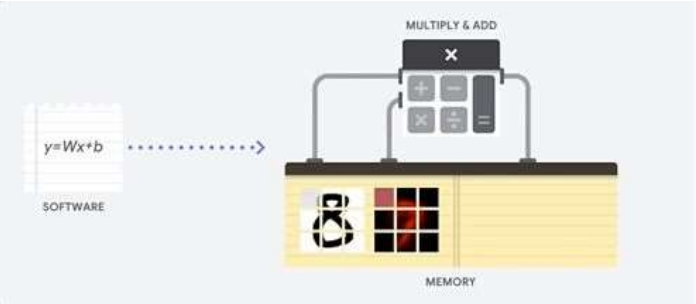
See our blog

Google Cloud

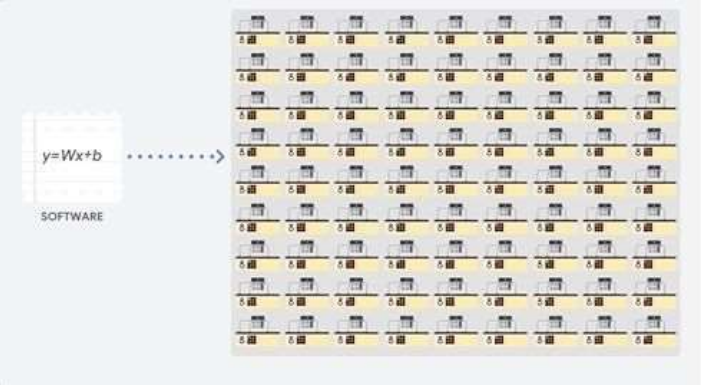
From CPU and GPU ...

Focus on Machine Learning and Deep Learning workloads

A CPU is a general-purpose processor based on the von Neumann architecture. That means a CPU works with software and memory like this.
The greatest benefit of CPUs is their flexibility. You can load any kind of software on a CPU for many different types of applications.



To gain higher throughput, GPUs contain thousands of Arithmetic Logic Units (ALUs) in a single processor. A modern GPU usually contains between 2,500–5,000 ALUs.
But, the GPU is still a general-purpose processor that has to support many different applications and software. Therefore, GPUs have the same problem as CPUs.

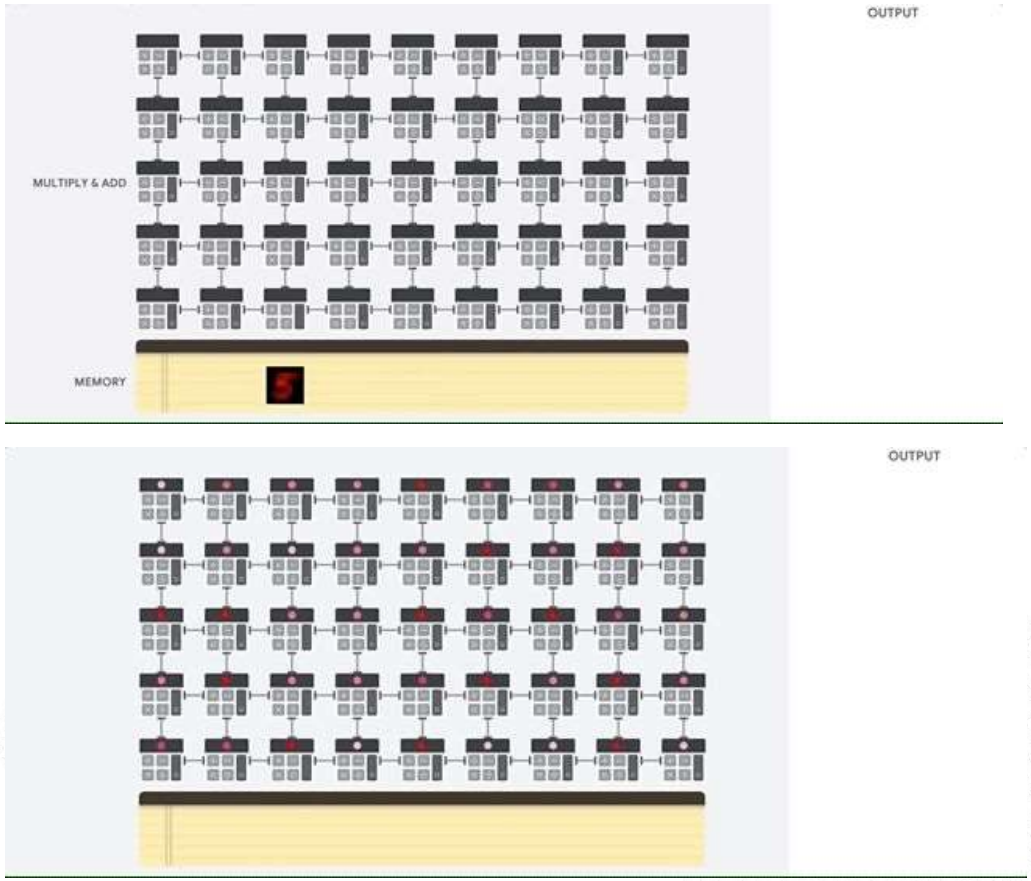


... to TPU

Focus on Machine Learning and Deep Learning workloads

Google designed Cloud TPUs as a matrix processor specialized for neural network workloads. TPUs can't run word processors, control rocket engines, or execute bank transactions, but they can handle massive matrix operations used in neural networks at fast speeds.

The primary task for TPUs is matrix processing, which is a combination of multiply and accumulate operations. TPUs contain thousands of multiply-accumulators that are directly connected to each other to form a large physical matrix. This is called a systolic array architecture.



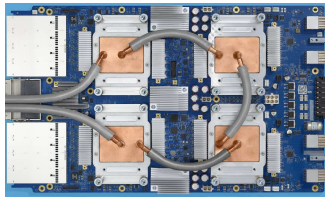
Google Cloud

TPU Devices: Across Generations

g.co/cloudtpu

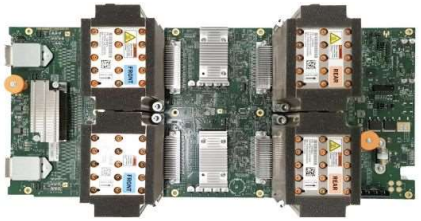
92 teraops (int8)
Inference only

TPU v1
(2015)



123 teraops (bf16)

Cloud TPU v3
(2018)



138 teraops (bf16)

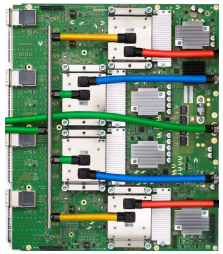
Cloud TPU v4 Lite
(2022)

Cloud TPU v2
(2017)

46 teraops (bf16)

Cloud TPU v4
(2021)

275 teraops (bf16)



9 exaflops
cluster (last
GoogleIO 2022)

More coming stay-tuned

Google

Identify the right technology (AI workloads)

When to choose CPU, GPU or TPU

CPUs

- Quick prototyping that requires maximum flexibility
- Simple models that do not take long to train
- Small models with small, effective batch sizes
- Models that contain many custom TensorFlow/PyTorch/JAX operations written in C++
- Models that are limited by available I/O or the networking bandwidth of the host system

GPUs

- Models with a significant number of custom TensorFlow/PyTorch/JAX operations that must run at least partially on CPUs
- Models with TensorFlow/PyTorch ops that are not available on Cloud TPU
- Medium-to-large models with larger effective batch sizes

TPUs

- Models dominated by matrix computations
- Models with no custom TensorFlow/PyTorch/JAX operations inside the main training loop
- Models that train for weeks or months
- Large models with large effective batch sizes

Reference models for Cloud TPUs

Machine translation & language modeling

Natural language processing:
- BERT
- Transformer
- Tensor2Tensor
- Mesh TensorFlow
- QANet
- Transformer-XL

Image recognition & object detection

Image recognition:
- AmoebaNet-D
- ResNet-50/101/152/200
- Inception v2/v3/v4
- MNasNet
- MobileNet

Object detection:
- RetinaNet
- DenseNet
- Mask R-CNN
- DeepLab

Speech recognition

ASR Transformer
Lingvo

Image generation

Image Transformer
DCGAN

github.com/tensorflow/tpu

Cloud TPU - Advantages

Speed Up Machine Learning Workloads

Designed from the ground up to accelerate machine learning & deep learning workloads

For both training and inferencing models

On-Demand Supercomputing

No upfront capital investment needed

Whether your task requires Cloud TPUs for hours or weeks, you can meet your AI workload needs without creating your own datacenter

Easy On-Ramp to Cloud

Because TensorFlow is open-source, you can take your existing ML workload that is already running in TensorFlow on Google Cloud TPU

Access Google's AI Innovation

Access the same accelerators that Google uses to empower Machine Learning in world-class Google products

Google Cloud

Problem
Visual search is a big part of the eBay experience. But delivering it is a major challenge; simply training their model with on-prem hardware took months.

Solution
That's when they switched to TPU Pods accessed through Google Cloud. We worked closely with eBay on a brand new image classifier designed specifically for TPUs. Next, we optimized those models by rapidly testing and fine-tuning them, also using TPU Pods. The result was an immediate **10%** boost in accuracy, and training time reduction of nearly **100X (here)**.

An image classifier trained on

Millions

product images

BILLION

product listings to search, visually

eBay used Cloud TPU Pods to make visual searcher faster and more accurate

+10%

increase in image recognition accuracy

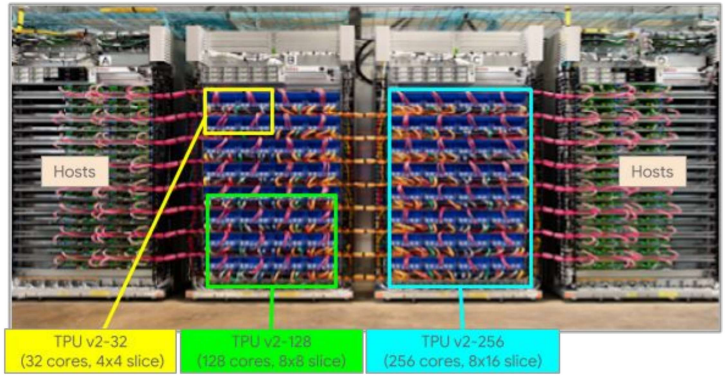
10×

Speedup in training time


Thank you

Cloud TPU Configurations

- TPU: The Tensor Processing Unit (TPU) is a custom-design chip, built from the ground up by Google for machine learning workloads.
- Cloud TPU: a device containing four TPU chips along with a fraction of a CPU host.
- Cloud TPU pods: Cloud TPUs are connected via a high-speed 2D toroidal mesh network to form Cloud TPU Pods.
- Cloud TPU slices: Slices, or smaller sections of pods, are scalable to address as much performance is needed for the workload. Slices are internal allocations consisting of different numbers of TPU cores. Pod slices come in 32, 128, 256, 512, 1024, and 2048 core-count configurations.



Google Cloud