Unleashing the Power of AI:

“Supercharge Your Vertex AI Workflows with the AI Hypercomputer (Cloud Next ‘24)”

Ananda Dwi Rahmawati
Google Developer Expert Cloud

Women Techmakers IWD 2024

Hello, GDG Bogor!

Ananda Dwi Rahmawati

- Cloud Engineer @ Activate Interactive Pte Ltd
- Google Developer Expert Cloud - Modern Architecture
- https://linktr.ee/misskecupbung

Introduction to Vertex AI

- A unified platform for building, deploying, and managing ML models on Google Cloud.
- Offers a range of services for all stages of the ML lifecycle, from data preparation to model deployment and monitoring.
- Designed to be scalable, secure, and easy to use, even for those with limited ML experience.
- Vertex AI has integrated with the following popular ML frameworks, such as: PyTorch, TensorFlow, scikit-learn, XGBoost

Vertex AI: Benefits and Key Feature

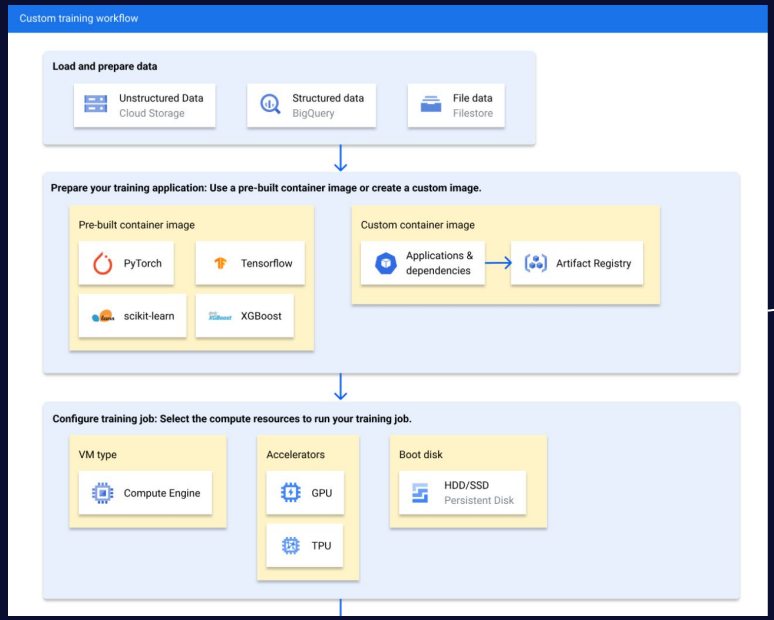
- **Fully Managed Infrastructure:** Train models without managing servers. Pay only for used resources. Vertex AI handles job logistics.
- **High Performance:** Optimized training surpassing GKE clusters. Identify and debug bottlenecks with TensorBoard Profiler.
- **Distributed Training:** Reduction Server accelerates multi-GPU training, reducing cost and time.
- **Hyperparameter Optimization:** Vertex AI tunes hyperparameters to find the best configuration for your model.

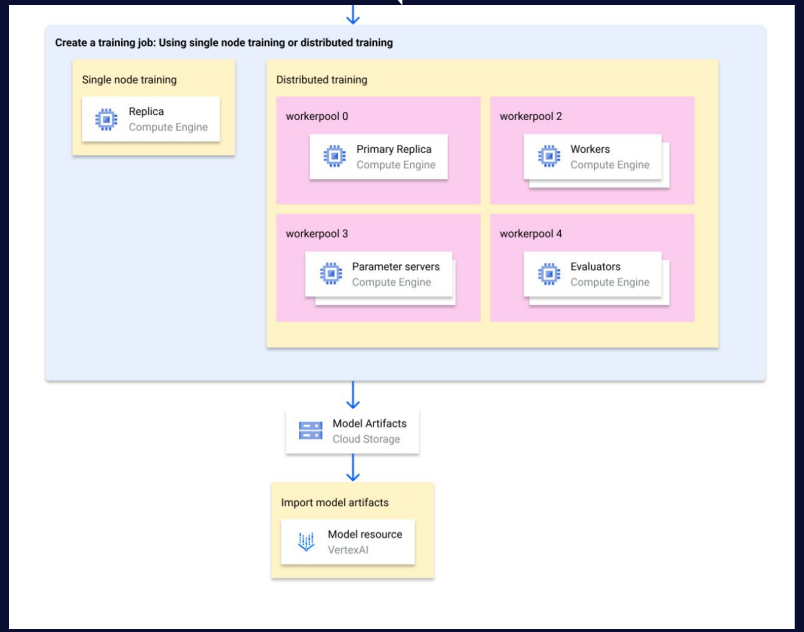
- **Enterprise Security:**
  - VPC peering restricts network access.
  - VPC Service Controls minimize data exfiltration risks.
  - Customer-managed encryption keys for compliance.
  - Identity and Access Management for granular control.
  - Single-tenant projects ensure data isolation.
- **Integrated MLOps Tools:**
  - Orchestrate ML workflows.
  - Perform feature engineering.
  - Run experiments.
  - Manage and iterate models.
  - Track ML metadata.
  - Monitor and evaluate model quality.

Vertex AI: Model training and deployment

- **AutoML**: Train models for various data types (tabular, image, text, video) without coding or data prep.
- **Custom Training:** Maintain full control, using preferred ML frameworks, custom code, and chosen hyperparameter tuning.
- **Model Garden:** Discover, test, customize, and deploy pre-built Vertex AI and open-source models and assets.
- **Generative AI:** Leverage Google's advanced generative AI models for text, code, images, and speech. Fine-tune them and deploy for your AI applications.

Vertex AI Custom Training





Vertex AI Custom Training

- Train custom models at scale
- JupyterLab with fully customizable compute
- Bring your ML code and run it in a Cloud with Vertex AI
- Keep track of your model experiments, use automated hyperparameter tuning, and leverage ML op's capabilities of Vertex AI
- Enterprise-grade

Vertex AI and AI Hypercomputer

- **Vertex AI acts as the interface**
  - To define, manage, and track your machine learning training jobs.
- **AI Hypercomputer provides the underlying infrastructure**
  - To provision and orchestrate the hardware resources needed for your training jobs. This includes accessing powerful TPUs and GPUs offered by GCP.

What is the AI Hypercomputer?

- The AI Hypercomputer is a next-generation computing platform specifically designed for demanding AI workloads.
- It boasts unparalleled processing power, massive memory capacity, and high-bandwidth networking capabilities.
- Optimized powerhouse architecture to handle complex AI models and expedite training and inference processes.

Update 1 - Advances in performance-optimized hardware

- Cloud TPU v5p GA: Google's most powerful and scalable TPU accelerator for training large generative AI models.
  - 2x the chips and higher FLOPS than TPU v4 pod.
  - Scales well with near-linear throughput improvement.
- GKE support for Cloud TPU v5p: Enables training and serving large AI models on GKE clusters.
  - TPU multi-host serving on GKE simplifies management of models across multiple hosts.
- A3 Mega with NVIDIA H100 GPUs: Coming next month, offering double the GPU-to-GPU networking bandwidth of A3. Confidential Computing on

Update 1 - Advances in performance-optimized hardware

- A3 VMs: Preview later this year, protects sensitive data and AI workloads.
- NVIDIA Blackwell GPUs on Google Cloud:
  - HGX B200 GPUs for demanding AI, data analytics, and HPC workloads (available now).
  - GB200 NVL72 GPUs (coming soon) for real-time LLM inference and massive-scale training.
- Customers using both Cloud TPUs and GPUs: Character.AI leverages both for efficient training and inference of large language models.

Update 2 - Storage optimized for AI/ML workloads

- **Cloud Storage FUSE caching (GA):** Improves training throughput by 2.9x and serving performance by 2.2x.
- **Parallelstore caching (preview):** Enables up to 3.9x faster training times and 3.7x higher throughput.
- **Filestore (GA):** Network file system for low latency data access, improving training times by up to 56%.
- **Hyperdisk ML (preview):** Next-generation block storage for AI inference, offering 12x faster model load times and high throughput.

Update 3 - Open software advancements

- **MaxText and MaxDiffusion reference implementations:** Open-source models for diffusion and large language models (LLMs) built on JAX.
- **JAX and OpenXLA optimizations:** Improved performance for JAX and OpenXLA on Cloud TPUs and GPUs.
- **PyTorch/XLA 2.3 support**: Brings new features for easier distributed training.
- **Optimum-TPU:** Performance-optimized package for training and serving Hugging Face models on TPUs.
- **Jetstream inference engine**: Open-source engine for high-throughput LLM inference on TPUs.
- **Open community models with NVIDIA**: Google models available as NVIDIA NIM inference microservices for flexible deployment.

Update 4 - New Dynamic Workload Scheduler modes

- Improves access and optimizes spend for AI workloads.
- Flex start mode (preview): Runs jobs ASAP based on availability, good for flexible start times. Now integrated with more services.
- Calendar mode (preview): Reserves AI computing capacity for up to 14 days (purchase 8 weeks in advance).
- Increased obtainability for GPUs and TPUs.
- Faster experiment iteration for researchers.

AI Hypercomputer Architecture

Flexible Consumption

Dynamic Workload Scheduler | On demand | CUD | Spot

Dynamic resource and cost optimization

Open software

PyTorch TensorFlow

Multislice Training, Multihost Inference, XLA

Google Kubernetes Engine & Compute

High flexibility, usability, and optionality

Performance-optimized hardware

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

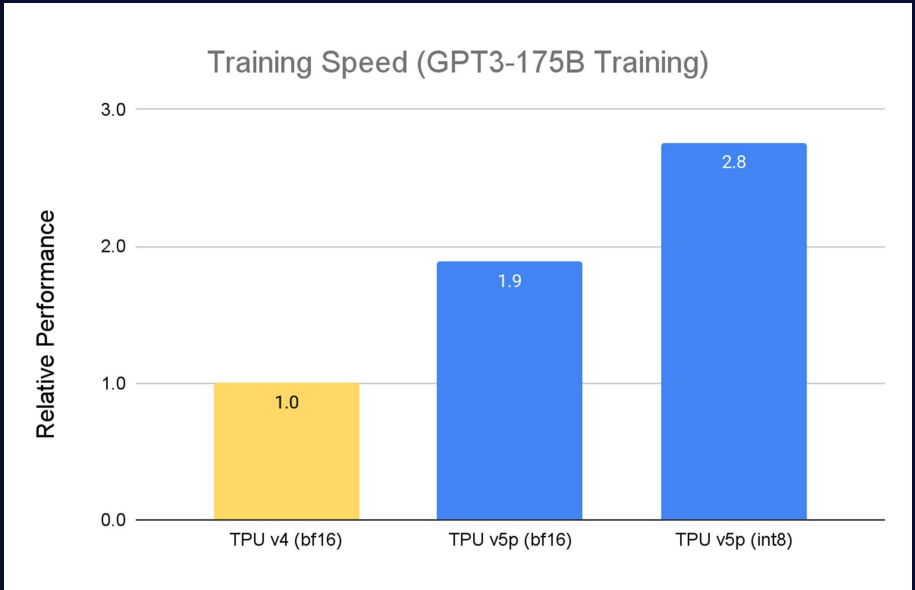
Exceptional performance and efficiency

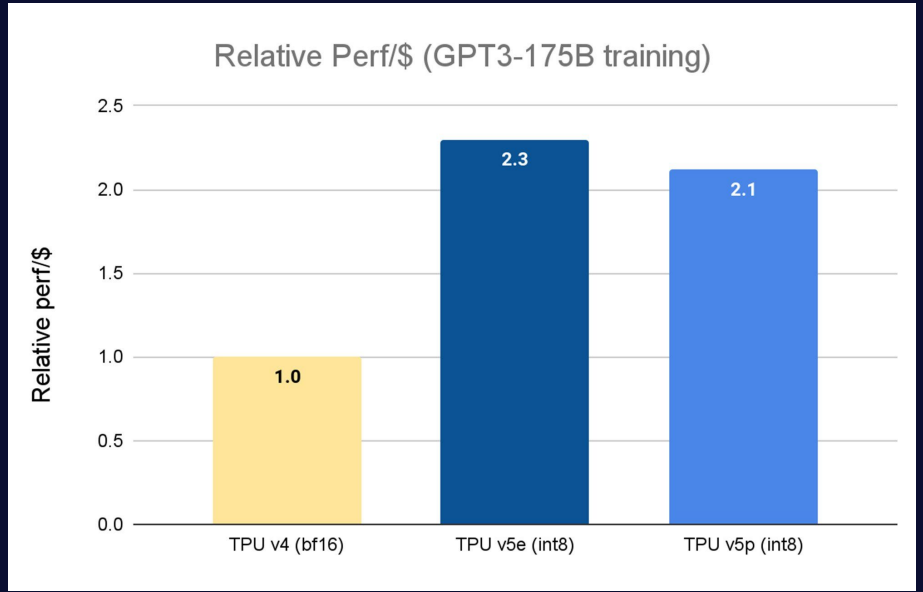
How to use/try AI Hypercomputer?

Google's AI Hypercomputer isn't directly accessible to users yet. It's an underlying architecture that powers some of Google Cloud's AI services, like Cloud TPUs and they are accessible through GCP. High-level overview of how GCP with TPUs can be used for AI workloads:

- **Identify your needs**: GCP offers different TPU configurations depending on your workload, e.g experimentation (short bursts), training (long periods), tuning (shorter durations), or serving (ongoing operations)?
- **Choose an interface**. Popular options include:
  - TensorFlow or PyTorch libraries with TPU support.
  - JAX, a high-performance numerical computation library from Google.
- **Set up your environment**: You'll need a GCP project and activate the TPU API.
- **Train or run your model:** Use your chosen interface to train your machine learning model or deploy it for inference using TPU acceleration.

How to use/try AI Hypercomputer?





	v4	v5e	v5p
Chips per pod	4096	256	8,960
Chip Bf16 TFLOPs	275	197	459
Chip Int8 TOPs	N/A	394	918
HBM (GB)	32	16	95
HBM BW (GB/s)	1228	820	2,765
ICI BW per chip (Gb/s)	2,400	1,600	4,800

TPU v5p is also 4X more scalable than TPU v4 in terms of total available FLOPs per pod.

References

- https://cloud.google.com/blog/products/compute/whats-new-with-google-clouds-ai-hypercomputer-architecture
- https://cloud.google.com/blog/products/ai-machine-learning/introducing-cloud-tpu-v5p-and-ai-hypercomputer
- https://cloud.google.com/blog/products/compute/announcing-cloud-tpu-v5e-and-a3-gpus-in-ga
- https://cloud.google.com/solutions/ai-hypercomputer?hl=en#ai-hypercomputer
- https://cloud.google.com/vertex-ai/docs/start/introduction-unified-platform
- https://cloud.google.com/vertex-ai/docs/training/overview


THANK YOU
GDG Bogor