Google Cloud

Next '24

How Anthropic uses Google Kubernetes Engine to run inference for Claude

Nova DasSarma

Cloud Infra Lead, Anthropic

Nathan Beach

Product Manager, Google Cloud

Ning Liao

Engineering Manager, Google Cloud

Agenda

01 Anthropic Serving Claude

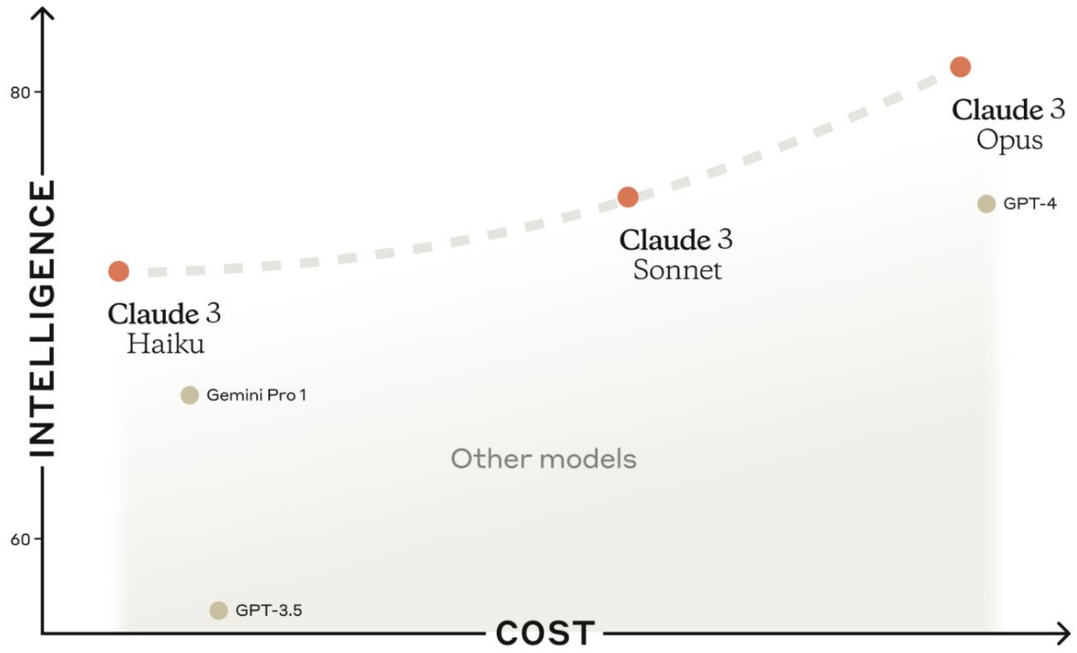
02 GKE Enables Large-Scale, Cost-Effective Inference

03 GKE Serving Gemma

Anthropic Serving Claude

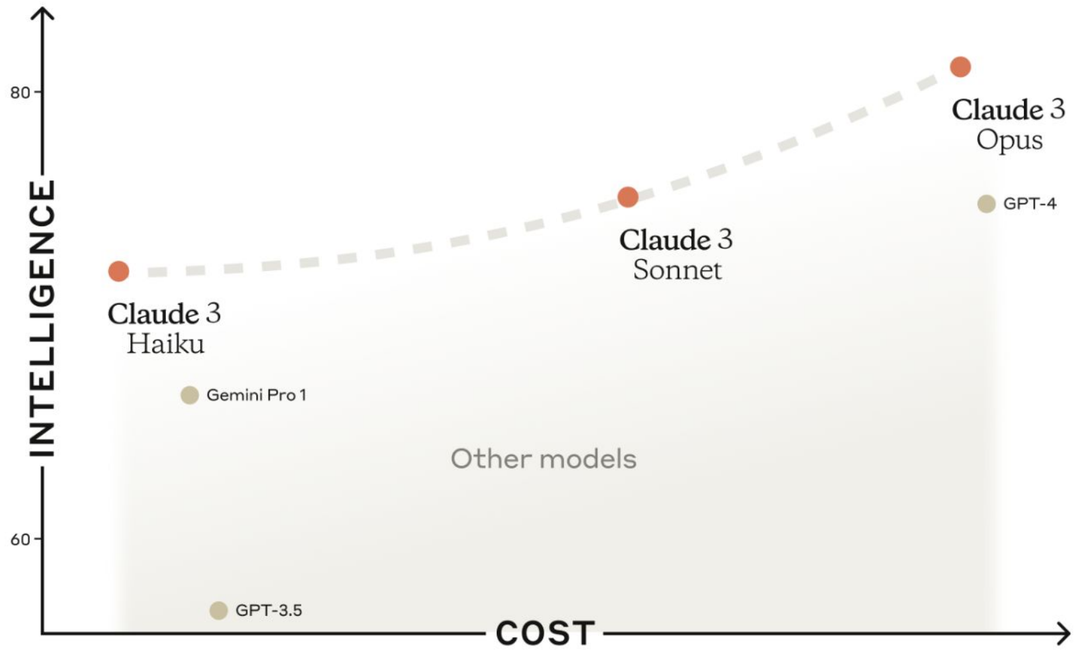
AWS is Anthropic's primary cloud provider

Anthropic



- Anthropic is an AI safety and research company founded in 2021, building reliable, interpretable, and steerable AI systems
- We develop large language models like the Claude 3 family, now available on Vertex

Anthropic: Mission



Our mission is to ensure AI helps people and society flourish by building frontier systems, studying their behaviors, working to responsibly deploy them, and regularly sharing our safety insights

Anthropic: Claude

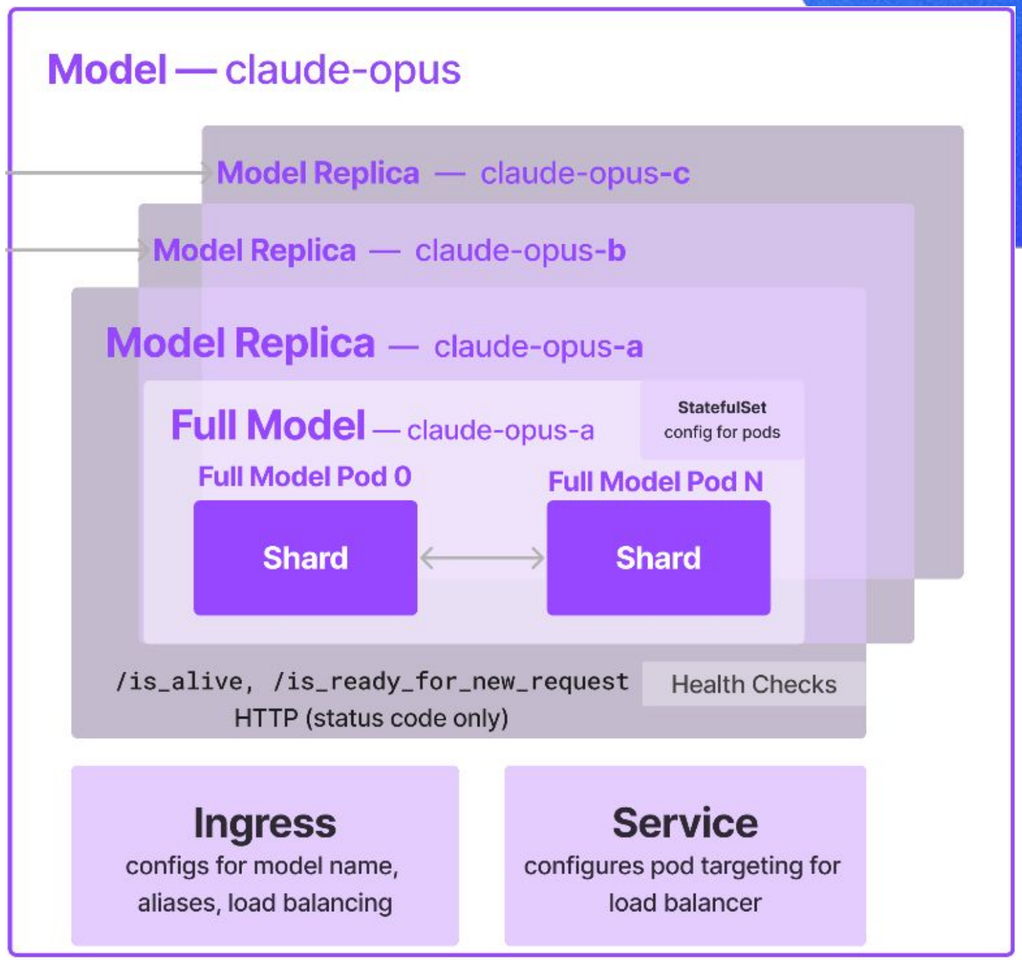
	Claude 3 Opus	Claude 3 Sonnet	Claude 3 Haiku	GPT-4	GPT-3.5	Gemini 1.0 Ultra
Undergraduate level knowledge *MMLU*	**86.8%** 5 shot	79.0% 5-shot	75.2% 5-shot	86.4% 5-shot	70.0% 5-shot	83.7% 5-shot
Graduate level reasoning *GPQA, Diamond*	**50.4%** 0-shot CoT	40.4% 0-shot CoT	33.3% 0-shot CoT	35.7% 0-shot CoT	28.1% 0-shot CoT	—
Grade school math *GSM8K*	**95.0%** 0-shot CoT	92.3% 0-shot CoT	88.9% 0-shot CoT	92.0% 5-shot CoT	57.1% 5-shot	94.4% Maj1@32
Math problem-solving *MATH*	**60.1%** 0-shot CoT	43.1% 0-shot CoT	38.9% 0-shot CoT	52.9% 4-shot	34.1% 4-shot	53.2% 4-shot
Multilingual math *MGSM*	**90.7%** 0-shot	83.5% 0-shot	75.1% 0-shot	74.5% 8-shot	—	79.0% 8-shot
Code *HumanEval*	**84.9%** 0-shot	73.0% 0-shot	75.9% 0-shot	67.0% 0-shot	48.1% 0-shot	74.4% 0-shot

The Claude 3 family

- Haiku - fast
- Sonnet - balanced
- **Opus - powerful**

MLOps at Scale

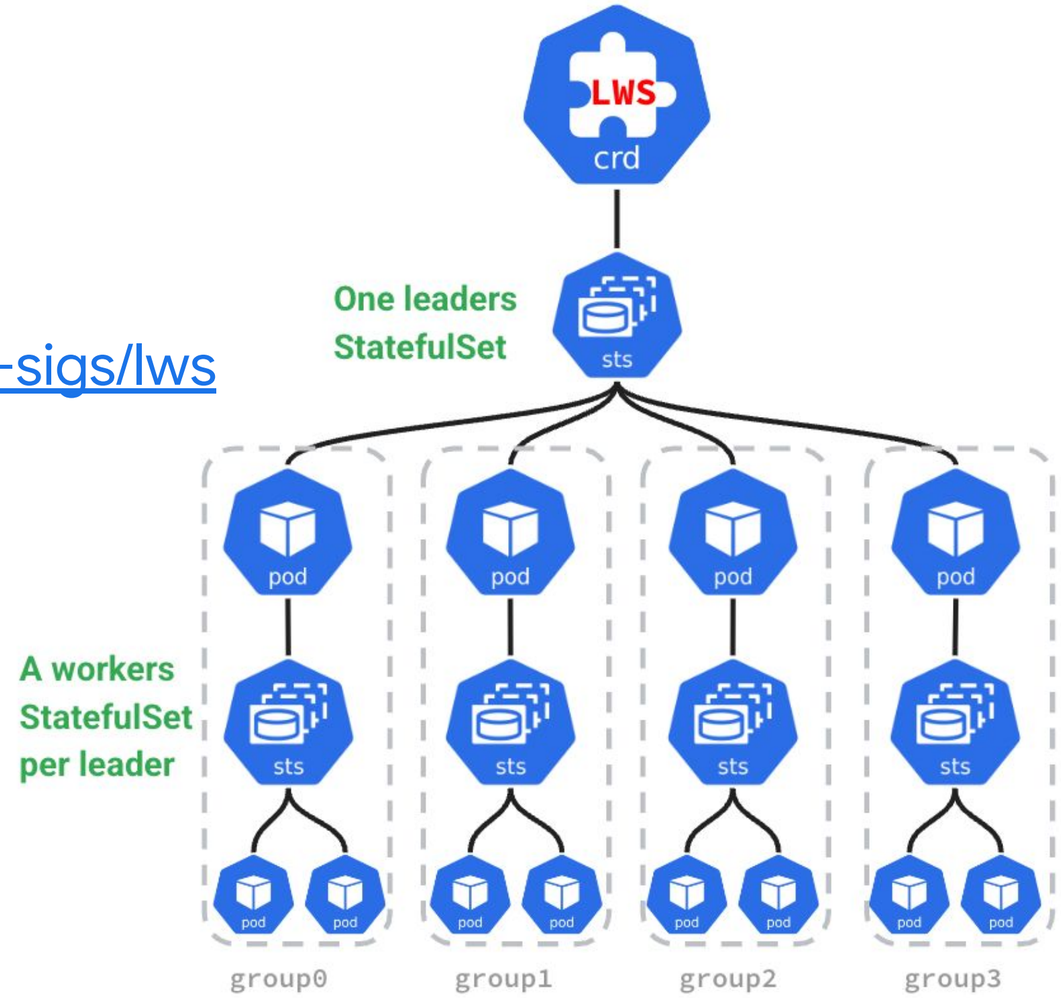
- Reliability challenges for individual servers
  - Host maintenance APIs
  - **`PodDisruptionBudgets`**
- `LeaderWorkerSet`



MLOps at Scale

- `LeaderWorkerSet`

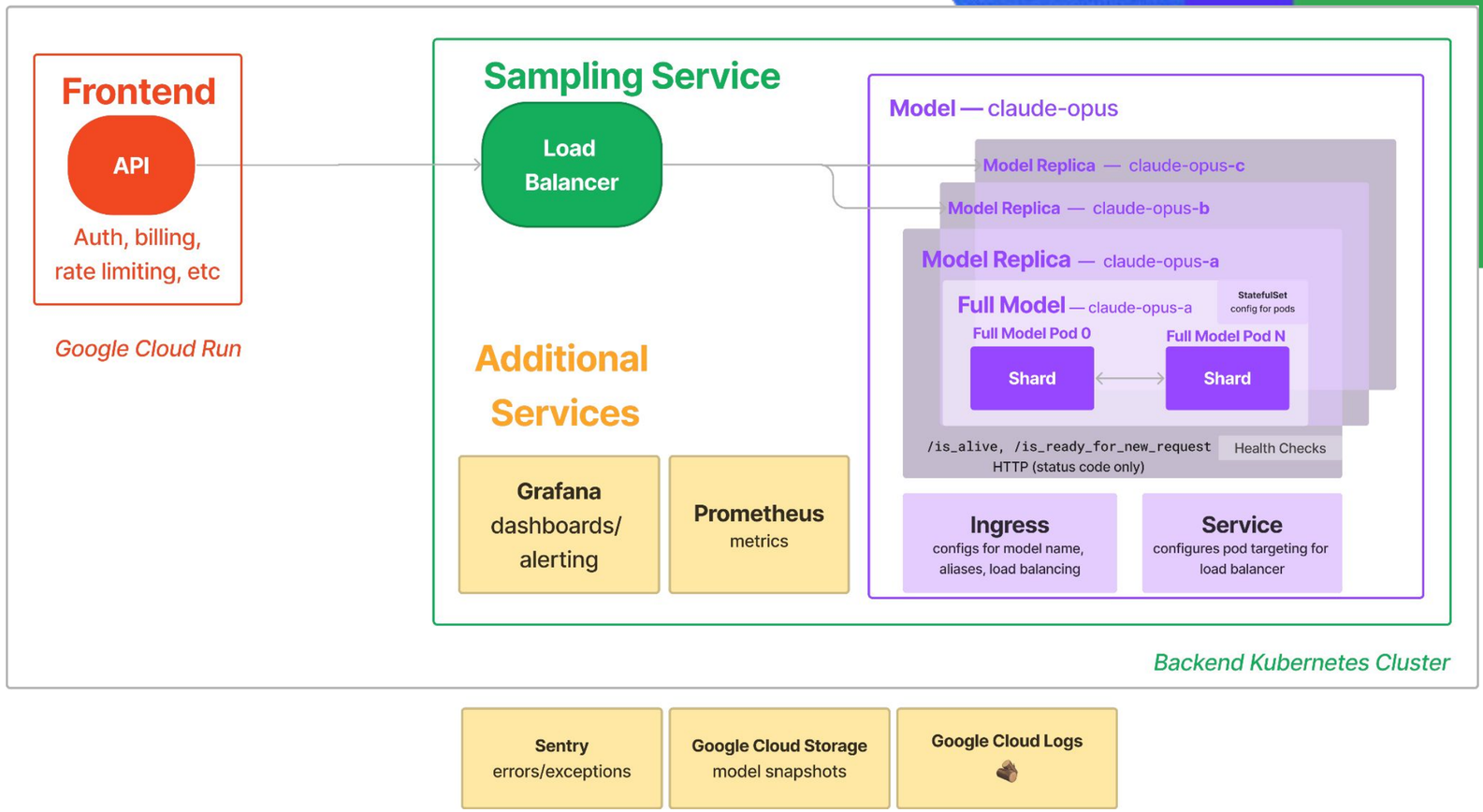
More at https://github.com/kubernetes-sigs/lws



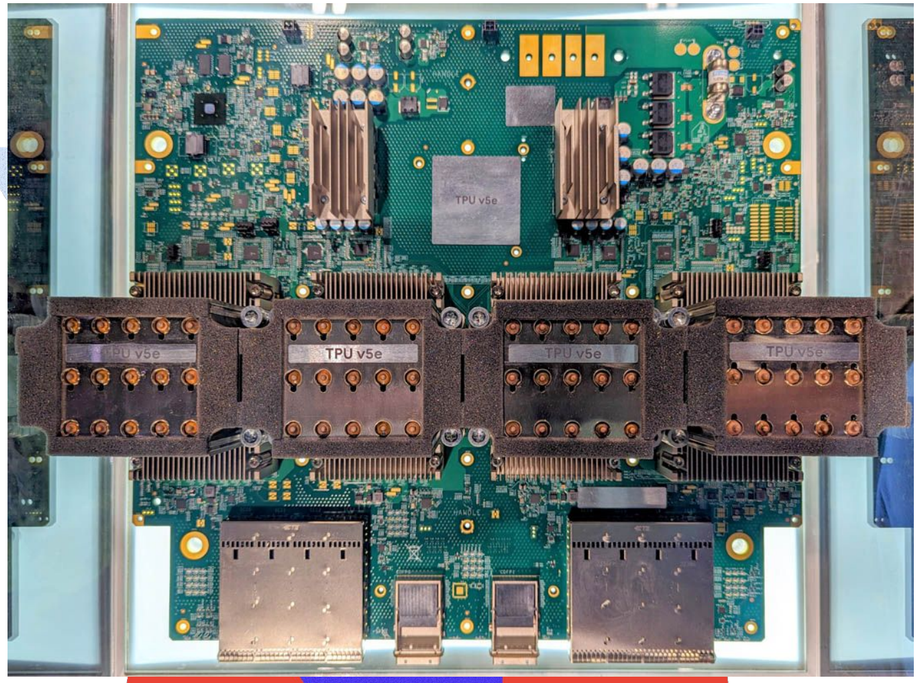
Kubernetes for Claude

- Kubernetes helps services communicate smoothly
- Optimizing utilization is easier with pods as an abstraction

  Pod Scheduling Readiness
- Multi-accelerator world



TPU v5e



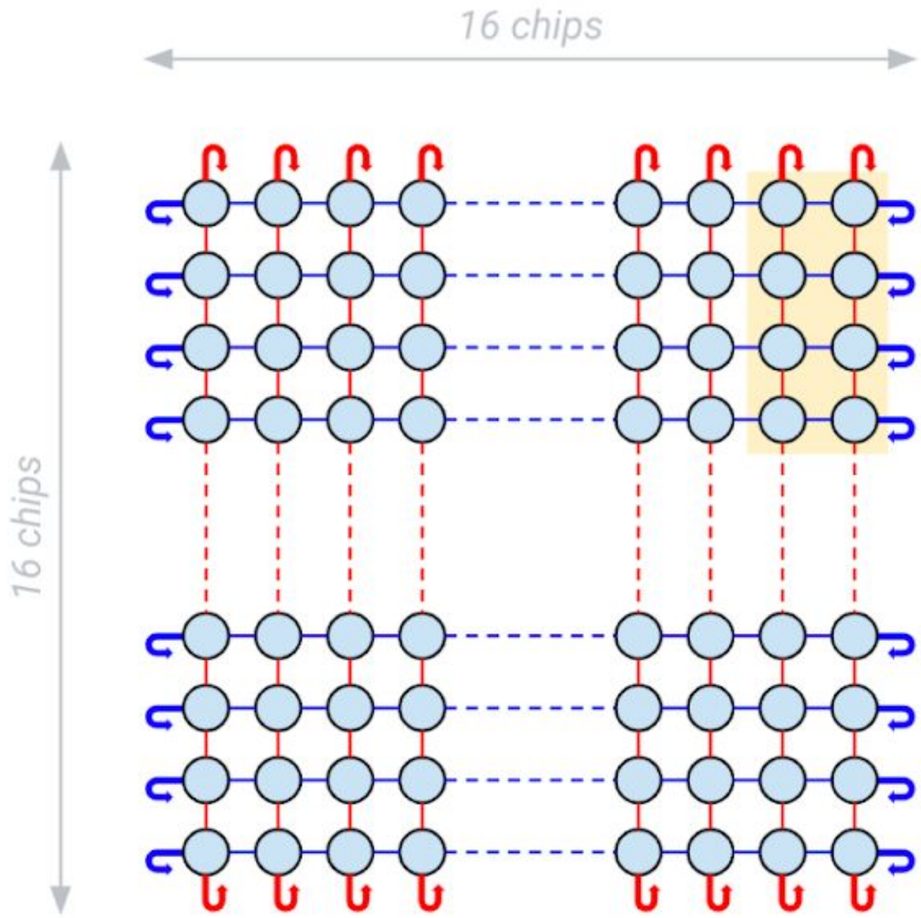
Each of these chips computes

393 teraops. per second.

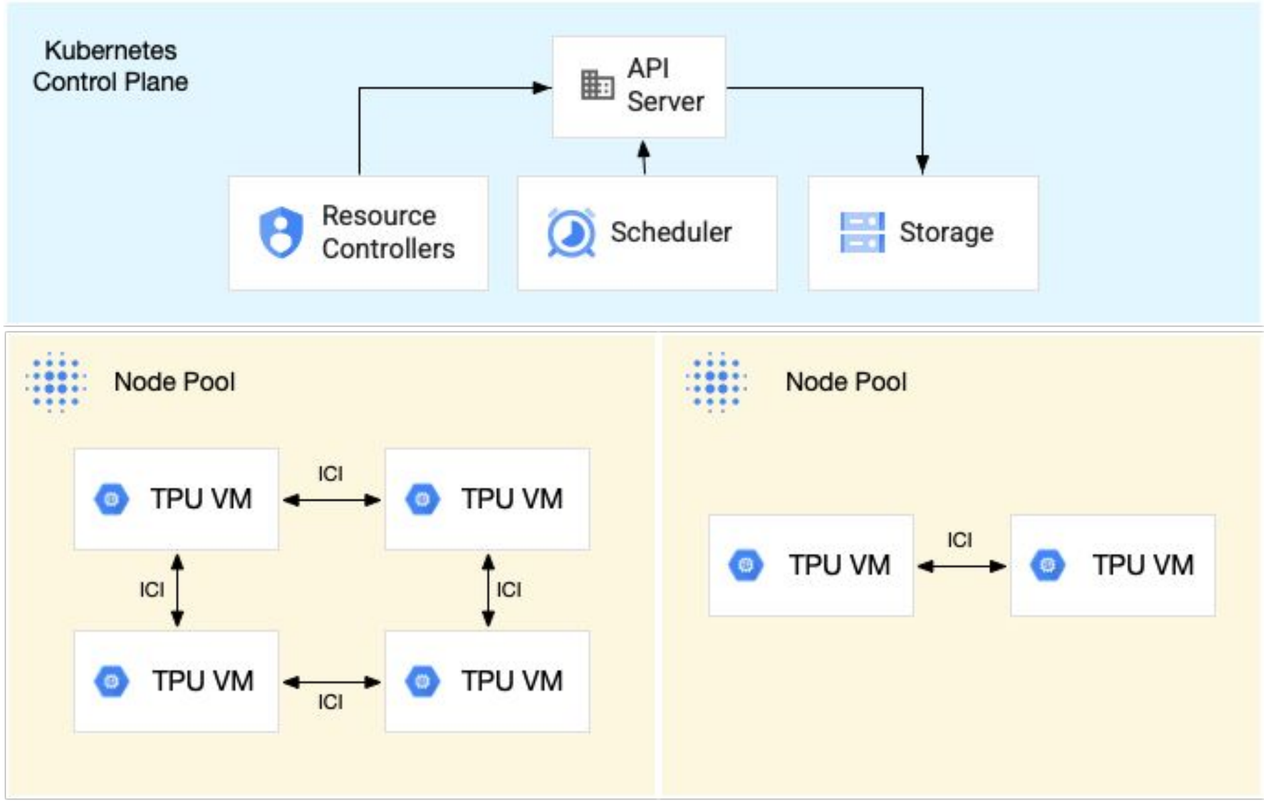
Int8 performance

Benefits of TPUs

- JAX speeds up model productionization
- ICI provides a performance boost for TPUs
- Cost-effective and scalable



GKE + TPU



ICI = inter chip connection

“

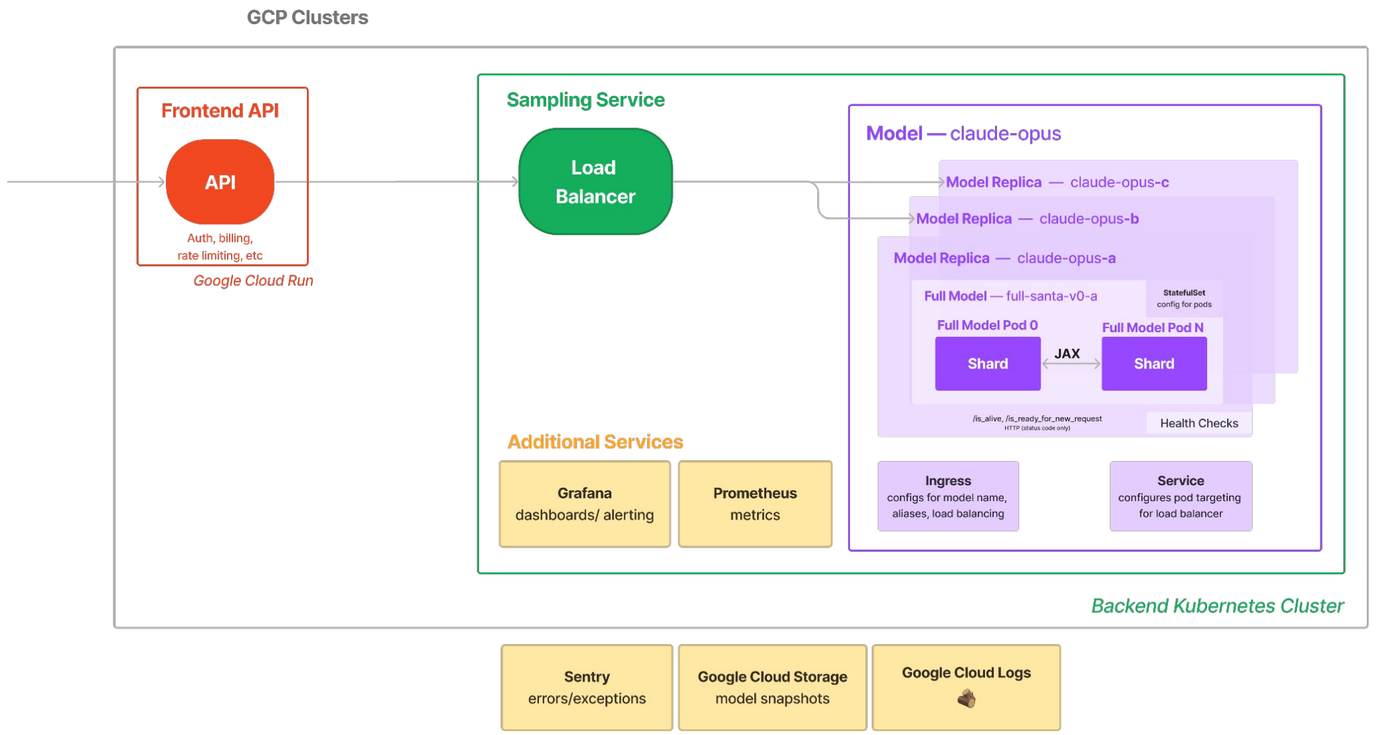
I find TPU optimization even more fun than GPU optimization. The architecture is simpler and more like a puzzle than a mystery box.”

Tristan Hume
Performance Lead

ANTHROP\C

Scalable Infrastructure, fast.

- Terraform for IaC makes managing capacity simple
- GKE gets our TPUs (and GPUs) serving customers *faster*

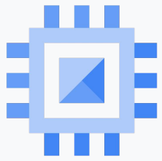


GKE Enables Large-Scale, Cost-Effective Inference

ML Frameworks & Libraries

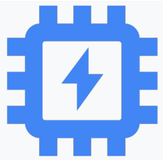
Model Server

Orchestration

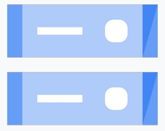


Compute

TPU



GPU



Storage

Networking

ML Frameworks & Libraries

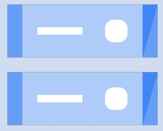
Model Server

Orchestration

Compute

TPU

GPU



Storage

Networking

TPU v5e for inference

Efficient

up to

2.7x

higher inference performance per dollar vs. TPU v4

Fast

up to

1.7x

latency speedup for LLMs vs. TPU v4

Scalable

Multihost

distributed serving supports large ML models

“Cloud TPU v5e consistently delivered up to 4X greater performance per dollar than comparable solutions in the market for running inference on our production ASR model.”

Domenic Donato
VP of Technology, AssemblyAI

G2 VMs with L4 GPUs

Fast

up to

4x

better performance than T4

Affordable

up to

40%

infrastructure cost savings relative to A10G

Global

Widespread

footprint across numerous regions

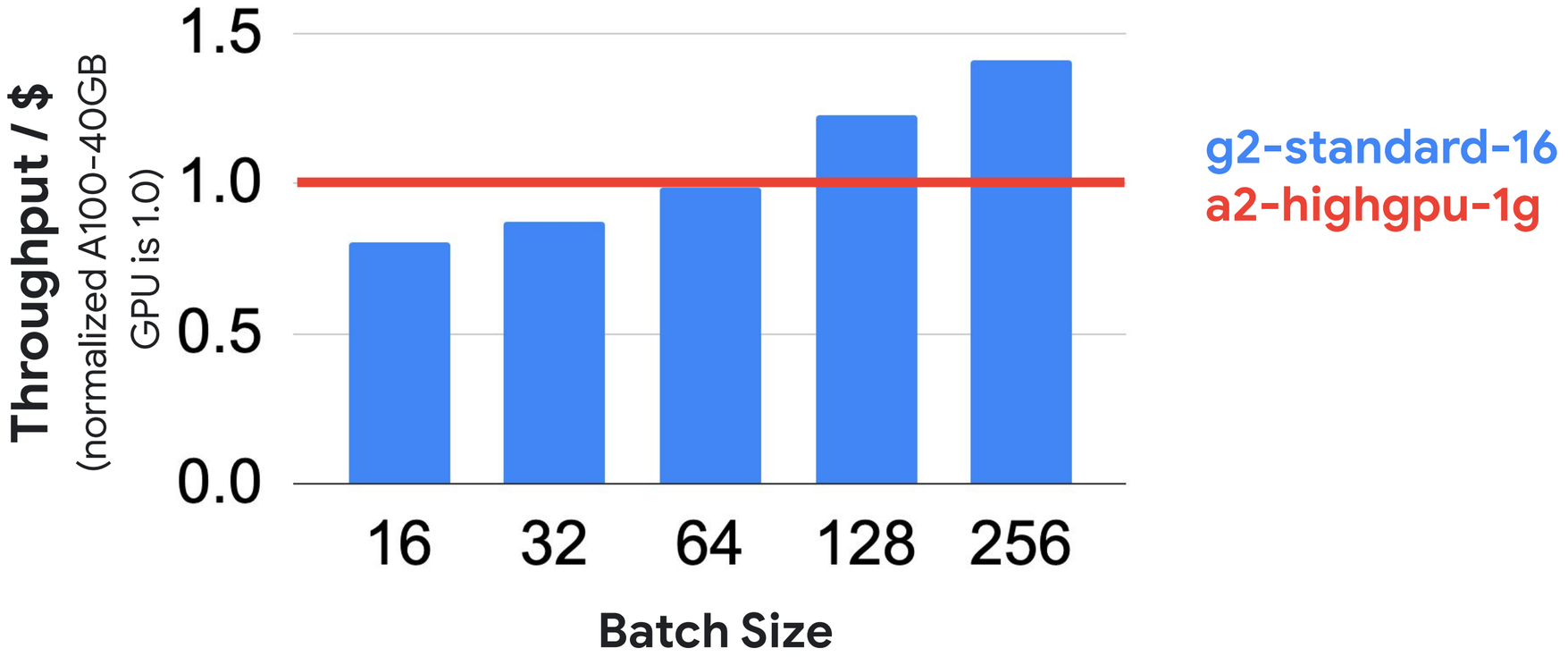
“

AppLovin is able to achieve nearly 2x improved price/performance compared with industry alternatives and support the company’s AI techniques.”

Omer Hasan
Vice President of Operations, AppLovin

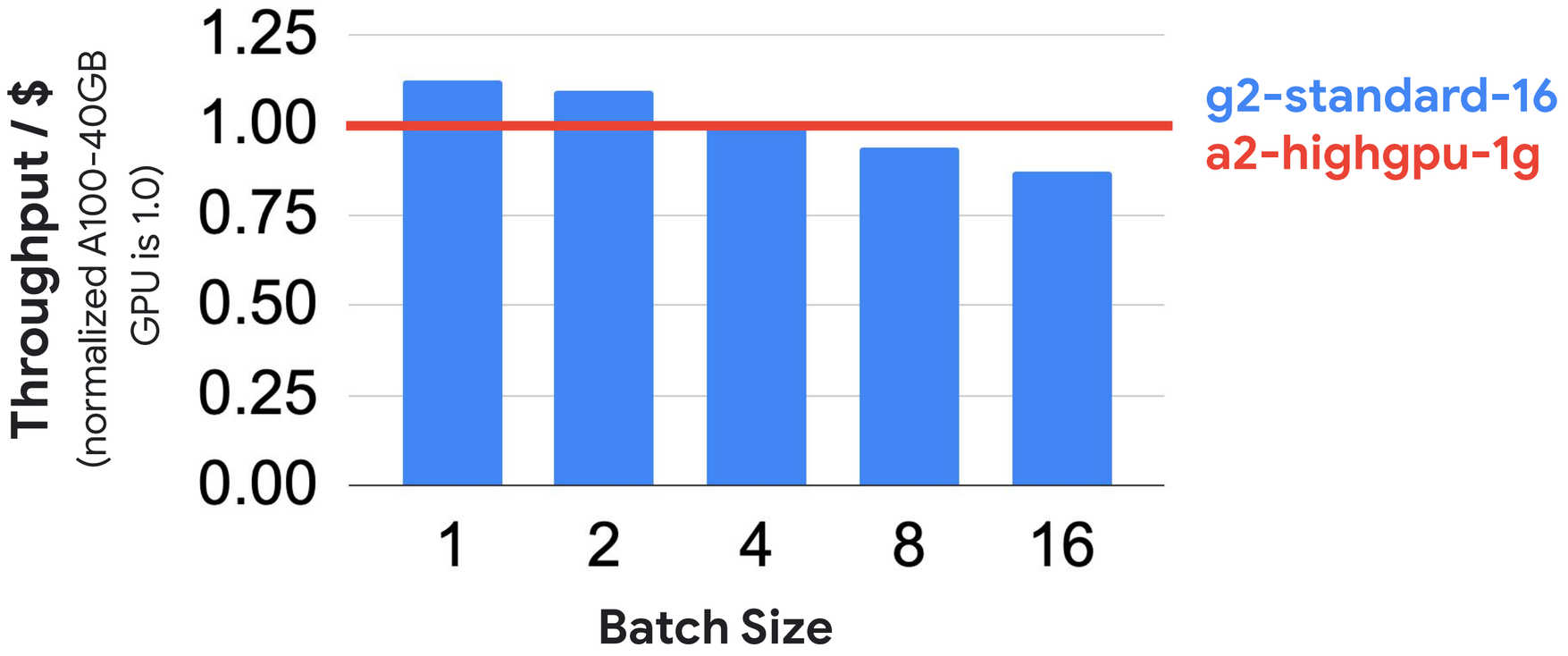
Optimize Throughput / $

When serving Llama 2 **7B**, which fits on a single L4 GPU, one **L4 GPU** provides better throughput per $ than one **A100-40GB GPU** at larger batch sizes.



Optimize Throughput / $

When serving Llama 2 **70B**, which requires many L4 GPUs, one **A100-40GB GPU** provides better throughput per $ than one **L4 GPU** at larger batch sizes.



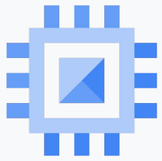
Recommendation

Among GPUs, choose L4 GPUs for smaller models and A100 GPUs for larger models.

ML Frameworks & Libraries

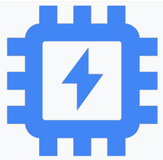
Model Server

Orchestration

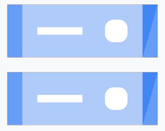


Compute

TPU



GPU



Storage

Networking

Fast Workload Startup

- **Pain point**: AI/ML container images can be very large (20GB+), making them very slow to load.
- **Solution**: Preload the container image on a secondary boot disk.
- **Also works** to cache data such as ML models, weights, etc.
- **Near-constant** latency even at massive scale.
- **In GKE**, enable with a single flag: `--secondary-boot-disk`

up to

29x

faster time to mount a 16GB container into *Running* status

“Within Vertex AI's prediction service, some of our container images can be quite large. After we enabled GKE container image preloading, our 16GB container images were pulled up to 29x faster in our tests.”

Shawn Ma
Software Engineer

Vertex AI

Cloud Storage FUSE

- **Pain point**: Reading data from Cloud Storage can add latency before a workload is ready to handle traffic.
- **Solution**: Data is streamed from Cloud Storage, allowing ML jobs to start much faster.
- **Also benefits** ML training thanks to read caching.
- **Portability** across clouds is an additional benefit; no code changes needed.

up to

3.3x

faster model load time

“

Using Cloud Storage Fuse with the GKE CSI driver has resulted not only in vastly simplified configuration for our applications, but has also reduced the pod startup time by up to 40%.”

Mark Chodos
Staff Engineer

GPU Time-Sharing

- **Pain point**: GPUs are very expensive, but most workloads don't fully utilize a GPU.
- **Solution**: Time-slice a single GPU so multiple containers can share it.
- **Perfect** for workloads with variable demand like inference and notebooks.
- **In GKE**, enable with a single flag: `gpu-sharing-strategy=time-sharing`

up to

66%

cost savings reported by customers

“We use GPU time-sharing between pods — increasing our average GPU utilization by 1.6x and lowering our costs by 66%.”

Ronald Griffin
Chief Technology Officer

Multi-Process Service

- **Pain point**: Time-sharing adds latency and overhead of full context switching between containers.
- **Solution**: NVIDIA Multi-Process Service (MPS) lets containers share a GPU in parallel with logical partitioning.
- **Perfect** for small to medium sized workloads, enabling greater GPU throughput.
- **In GKE**, enable with a single flag: `gpu-sharing-strategy=mps`

GPU & TPU on Autopilot

- **Pain point**: Managing infrastructure takes time away from growing your business.
- **Solution**: Autopilot supports all the latest GPUs (H100, A100, L4, T4) and TPUs (v4, v5e, and v5p).
- **Solution**: Autopilot supports existing GCE reservations and committed use discounts (CUDs) for GPUs and TPUs.
- **In GKE**, add a label to your workload such as: `cloud.google.com/gke-accelerator: nvidia-h100-80gb`

SLA

at the Pod level, backed by Google SREs

“With GKE Autopilot, we can easily scale our pods, optimize our resource utilization, and ensure the security and availability of our nodes. We are excited to use GKE Autopilot to power Contextual Language Models while saving us money and improving our performance.”

Soumitr Pandey
Member of Technical Staff

GKE Serving Gemma

Kubernetes is the foundation

GKE is the most scalable leading Kubernetes service available in the industry today.

Accelerator	Framework	Inference Stack
Cloud GPU	PyTorch	vLLM, Hugging Face TGI, Triton + TensorRT-LLM
Cloud TPU	JAX PyTorch	JetStream

Gemma on GKE

Integration

- Hugging Face
- Vertex Model Garden
- Google Colab Enterprise Notebook
- Kaggle

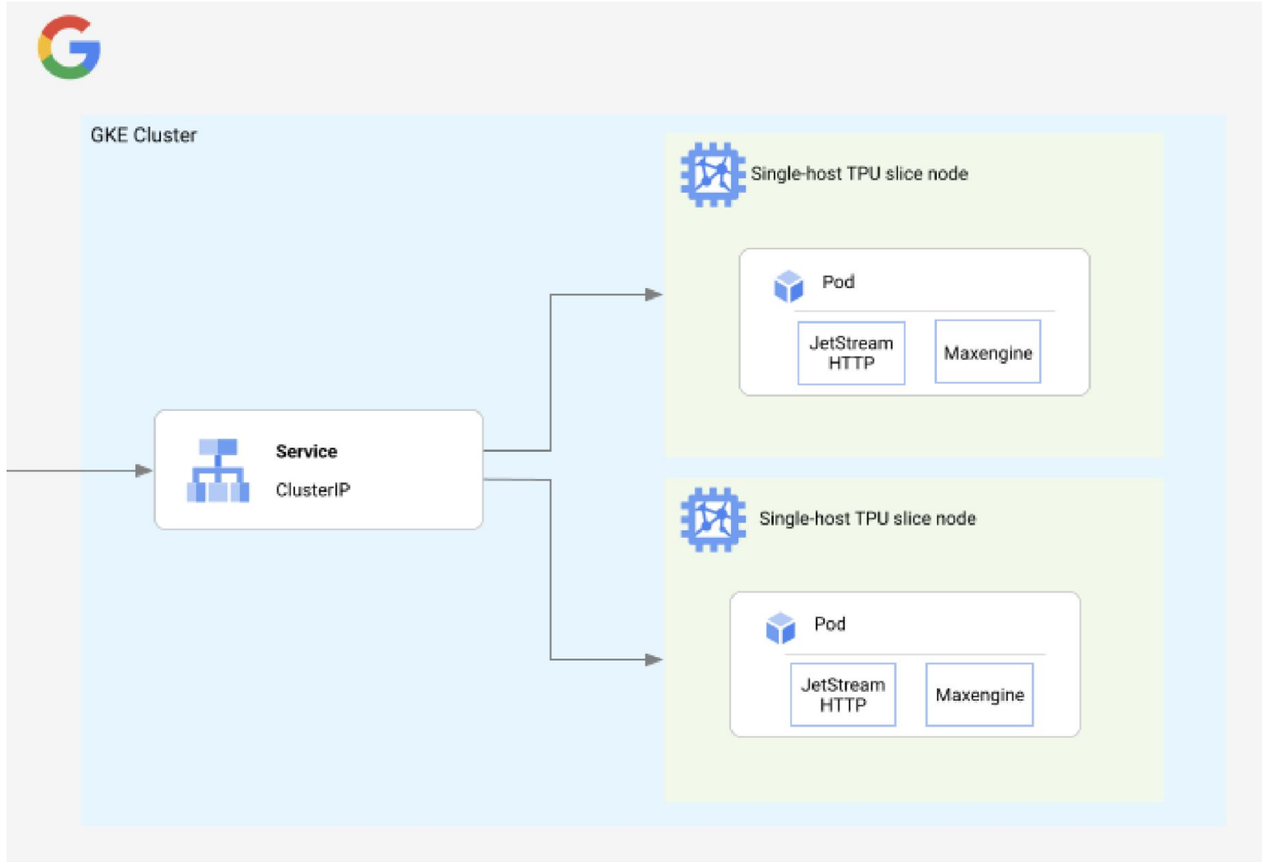
Serving Stack

- JetStream
- Nvidia TensorRT-LLM & NIM
- vLLM
- Hugging Face TGI

JetStream on GKE

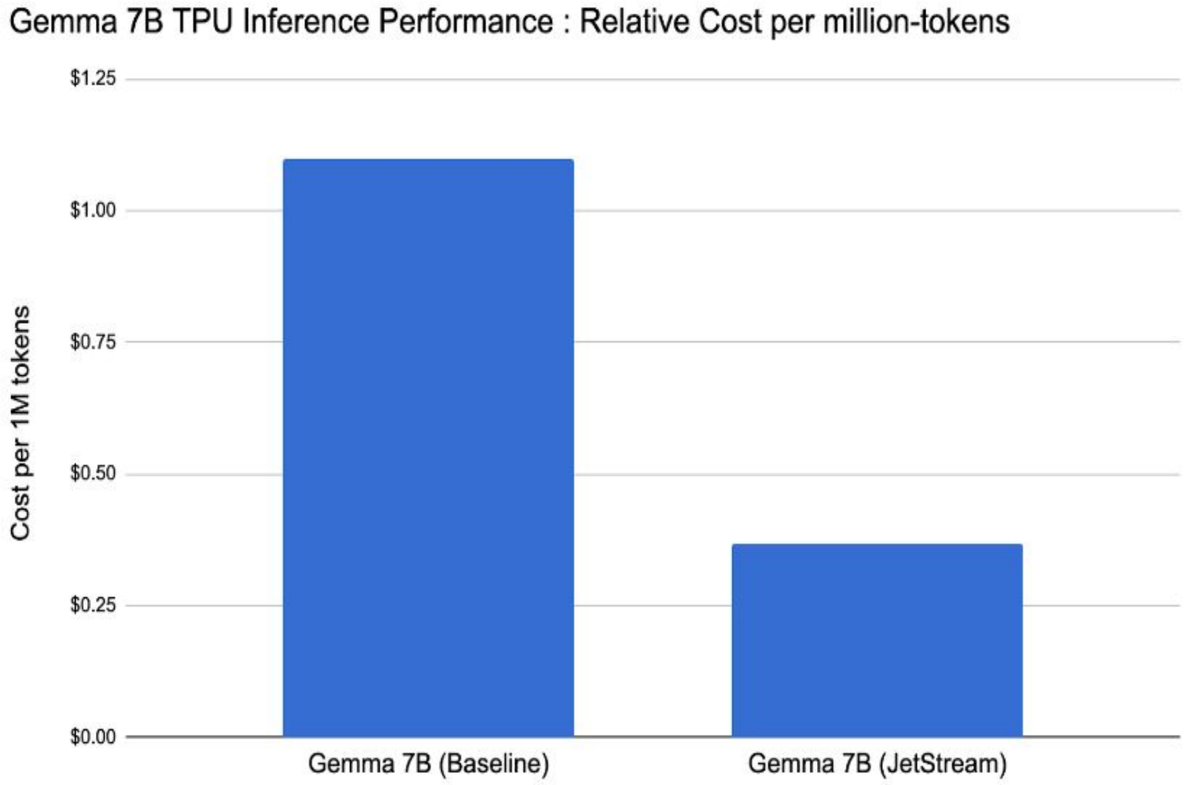
- An open-source, high-performance, cost-efficient LLM Inference for JAX and PyTorch/XLA.

  https://github.com/google/JetStream



https://cloud.google.com/kubernetes-engine/docs/tutorials/serve-gemma-tpu-jetstream

Gemma+JetStream on GKE



Google internal data. Measured using Gemma 7B (MaxText) on TPU v5e-8. Input length 1024, output length 1024 for a specific request rate and batch size. Continuous batching, int8 quantization for weights, activations, KV cache. As of April, 2024.

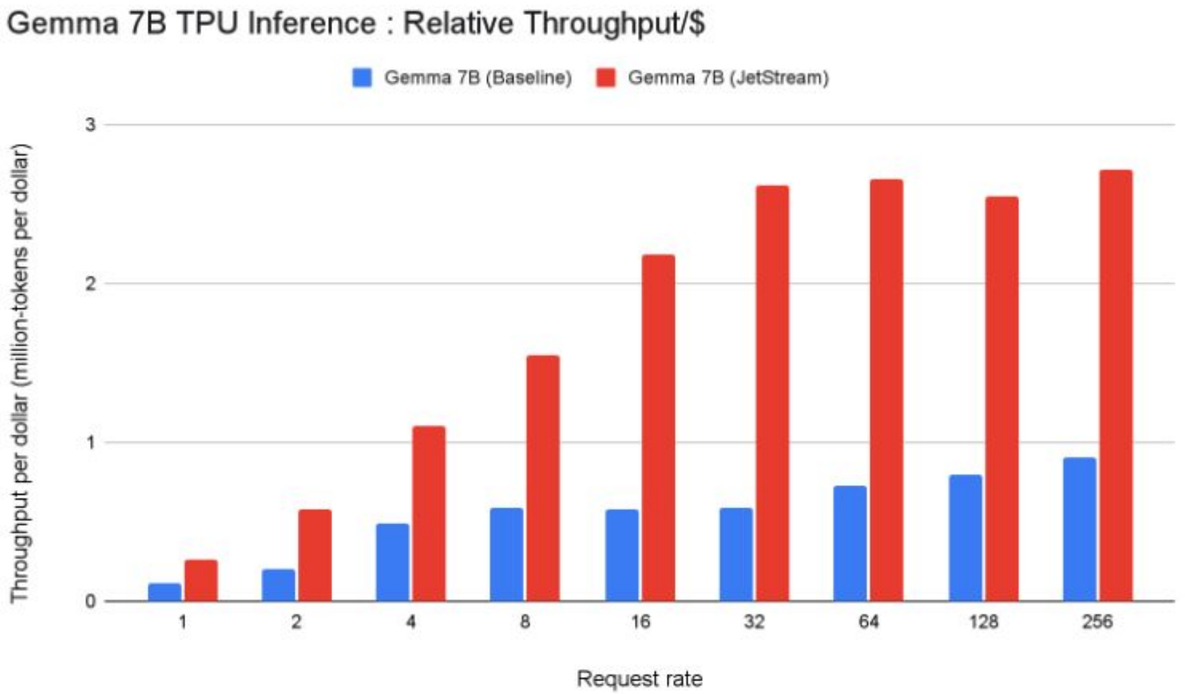
up to

~3x

Reduction in cost per 1M tokens

Gemma + JetStream on GKE

- Throughput per dollar for serving on JetStream is consistently higher than baseline, even for higher request rates.



Google internal data. Measured using Gemma 7B (MaxText) on TPU v5e-8. Input length 1024, output length 1024 for varying request rate from 1 to 256. Continuous batching, int8 quantization for weights, activations, KV cache. As of April, 2024.

GKE AI Benchmarking

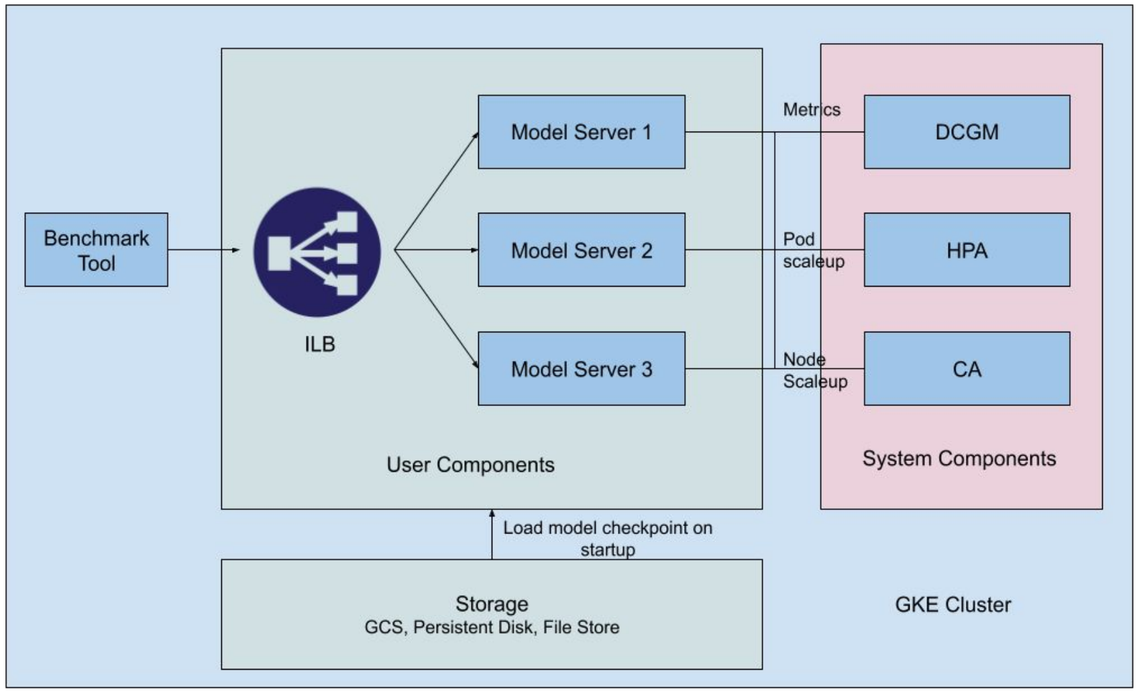
Blogs

https://cloud.google.com/blog/products/containers-kubernetes/serving-gemma-on-google-kubernetes-engine-deep-dive

https://cloud.google.com/blog/products/ai-machine-learning/performance-deepdive-of-gemma-on-google-cloud

Try it out!

https://github.com/GoogleCloudPlatform/ai-on-gke/tree/main/benchmarks



We are interested in your feedback!

Connect with a GKE/Serverless PM or UX researcher.

kubernetes turns 10!

#k8sturns10

Ready to build what's next?

Tap into **special offers** designed to help you **implement what you learned** at Google Cloud Next.

Scan the code to receive personalized guidance from one of our experts.

Or visit **g.co/next/24offers**

Upcoming Sessions

1

Tomorrow at 10:15am
OPS209
From RAG to autonomous apps with Weaviate and Gemini on Google Kubernetes Engine

2

Tomorrow at 11:30am
OPS302
Maximize machine learning productivity at scale

3

Tomorrow at 12:15pm
DEV309
Run large-scale AI training and inference for Llama 2 on Cloud Accelerators


Thank you