# Google Kubernetes Engine (GKE) boosted AI inferencing compared to Amazon EKS

*Principled Technologies found GKE with GKE Inference Gateway delivered 15.7% higher token throughput, 92.8% lower latency, and significantly lower tail latency.*

SAN JOSE, CA, UNITED STATES, May 22, 2026 /EINPresswire.com/ -- As more organizations deploy generative AI applications, infrastructure performance can play a critical role in serving model responses quickly and efficiently. A new hands-on performance report from Principled Technologies (PT) shows that an inference engine running in Google Kubernetes Engine (GKE) with GKE Inference Gateway outperformed the same engine running in Amazon Elastic Kubernetes Service (EKS) using a standard HTTP load balancer for the Llama 3.1-8B Instruct model on identical hardware. The PT evaluation used the Kubernetes inference-perf benchmark on inference-engine deployments backed by eight NVIDIA A100 40GB GPUs.

Accelerate your generative AI inferencing by choosing Google Kubernetes Engine

Key takeaways

The PT study found meaningful improvements across throughput, latency, and stability:

- 15.7% higher output token throughput—The GKE solution processed roughly 1,000 more tokens per second than the Amazon EKS solution, enabling greater capacity or reduced hardware needs for equivalent workloads.

- 92.8% lower time to first token (TTFT)—GKE delivered a mean TTFT more than 2,000 milliseconds lower than Amazon EKS, which could dramatically improve perceived responsiveness for interactive AI applications.
- 62.6% lower inter-token latency (ITL)—Mean ITL on GKE was lower compared to Amazon EKS, potentially yielding smoother streaming and faster token emission after the initial response.
- Significantly improved tail latency and stability—GKE showed up to 83.9% lower 95th-percentile tail latency and a 67.0% lower 95th-percentile normalized time per output token, which could reduce the incidence of extremely slow responses under load.

The report attributes these gains to inference-aware optimizations provided by the GKE Inference Gateway, including prefix-cache-aware routing, which directs requests with shared context to the same model replica to maximize cache hits. These capabilities can reduce redundant computation, better use GPU and TPU accelerators, and improve both throughput and latency—benefits particularly relevant to multi-turn AI chat, retrieval-augmented generation (RAG), and document Q&A scenarios where requests commonly share prefixes or context.

The PT report states, “Companies that rely on workloads where requests commonly share prefixes or benefit from cache locality (for example, document Q&A, multi turn conversations, or template-based generation) need high performance. For these workloads, consider GKE with GKE Inference Gateway to improve responsiveness, capacity, and cost efficiency on equivalent GPU hardware.”

FAQ

Who conducted this evaluation?
A: Principled Technologies (PT) performed the hands-on performance evaluation.

What was tested?
A: PT compared the inference performance of the Llama 3.1-8B Instruct model on two cloud environments that differed only in how they distribute requests to multiple engines. The first environment was Google Kubernetes Engine (GKE) with GKE Inference Gateway, and the second environment was Amazon Elastic Kubernetes Service (EKS) with a standard HTTP load balancer.

What hardware and configurations did PT use?
A: Both cloud solutions were backed by eight NVIDIA A100 40GB GPUs; the primary difference between the solutions was GKE using the inference-aware GKE Inference Gateway versus Amazon EKS using a standard HTTP load balancer.

What key performance improvements did PT observe?
A: PT measured 15.7% higher token throughput, 92.8% lower time to first token (TTFT), 62.6% lower inter-token latency (ITL), and up to 83.9% lower 95th-percentile tail latency for GKE vs Amazon EKS.

Why did GKE perform better?
A: The report attributes gains to inference-aware optimizations in the GKE Inference Gateway.

Which workloads can benefit most from these gains?
A: Interactive generative AI workloads—multi-turn chat, streaming interfaces, retrieval-augmented generation (RAG), and document Q&A—are especially likely to see improved responsiveness and infrastructure efficiency.

About the report

PT performed the analysis, including methodology and metric definitions (TTFT, TPOT, ITL, NTPOT, and tail latency). PT used cloud specific vLLM tuning sets and the Kubernetes inference-perf tool to capture throughput and latency behavior across varying request rates.

Learn more about the results from PT testing and what they could mean for organizations seeking to run AI in the cloud.

About Principled Technologies, Inc.
Principled Technologies, Inc. is the leading provider of technology marketing and learning & development services.

Principled Technologies, Inc. is located in Durham, North Carolina, USA. For more information, please visit www.principledtechnologies.com.

Sharon Horton
Principled Technologies, Inc.
press@principledtechnologies.com
Visit us on social media:
LinkedIn
Facebook
YouTube
X

---