

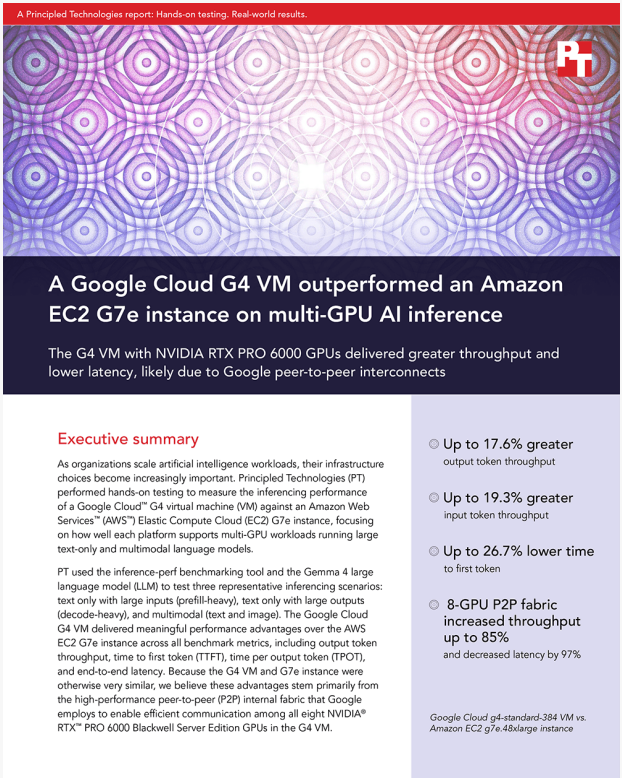
Google Cloud G4 VM outperformed Amazon EC2 G7e on multi-GPU AI inference in Principled Technologies study

Hands-on benchmark testing revealed throughput and latency advantages, likely due to the all-to-all P2P GPU fabric in the Google G4 VM

SAN JOSE, CA, UNITED STATES, October 1, 2026 /EINPresswire.com/ --

Principled Technologies (PT), a fact-based technology research and testing firm, has released results of a hands-on comparison of a Google Cloud G4 virtual machine against an Amazon Web Services (AWS) Elastic Compute Cloud (EC2) G7e instance on multi-GPU AI inference. Both platforms were configured with eight NVIDIA RTX PRO 6000 Blackwell Server Edition GPUs, but PT testing found the Google Cloud G4 VM delivered superior performance in two of the three test scenarios, with advantages up to 17.6% in output token throughput, 19.3% in input token throughput, and 26.7% in time to first token.

PT used the inference-perf benchmarking tool—developed by the Kubernetes SIG community—to measure performance of the Google Gemma 4 large language model across three inference scenarios: multimodal (text plus image input), prefill-heavy text (large input, modest output), and decode-heavy text (modest input, large output). Each scenario represents a different pattern of multi-GPU workload demand.



Principled Technologies®

A Principled Technologies report: Hands-on testing. Real-world results.

A Google Cloud G4 VM outperformed an Amazon EC2 G7e instance on multi-GPU AI inference

The G4 VM with NVIDIA RTX PRO 6000 GPUs delivered greater throughput and lower latency, likely due to Google peer-to-peer interconnects

Executive summary

As organizations scale artificial intelligence workloads, their infrastructure choices become increasingly important. Principled Technologies (PT) performed hands-on testing to measure the inferencing performance of a Google Cloud™ G4 virtual machine (VM) against an Amazon Web Services™ (AWS™) Elastic Compute Cloud (EC2) G7e instance, focusing on how well each platform supports multi-GPU workloads running large text-only and multimodal language models.

PT used the inference-perf benchmarking tool and the Gemma 4 large language model (LLM) to test three representative inferencing scenarios: text only with large inputs (prefill-heavy), text only with large outputs (decode-heavy), and multimodal (text and image). The Google Cloud G4 VM delivered meaningful performance advantages over the AWS EC2 G7e instance across all benchmark metrics, including output token throughput, time to first token (TTFT), time per output token (TPOT), and end-to-end latency. Because the G4 VM and G7e instance were otherwise very similar, we believe these advantages stem primarily from the high-performance peer-to-peer (P2P) internal fabric that Google employs to enable efficient communication among all eight NVIDIA® RTX™ PRO 6000 Blackwell Server Edition GPUs in the G4 VM.

- Up to 17.6% greater output token throughput
- Up to 19.3% greater input token throughput
- Up to 26.7% lower time to first token
- 8-GPU P2P fabric increased throughput up to 85% and decreased latency by 97%

Google Cloud g4-standard-384 VM vs. Amazon EC2 g7e.48xlarge instance

A Google Cloud G4 VM outperformed an Amazon EC2 G7e instance on multi-GPU AI inference September 2026

Report: A Google Cloud G4 VM outperformed an Amazon EC2 G7e instance on multi-GPU AI inference

Key findings

In the multimodal scenario, the Google G4 VM outperformed the Amazon G7e instance across every metric PT measured: 17.6% greater output token throughput (284.3 vs. 241.8 tokens per second), 19.3% greater input token throughput (2,775.2 vs. 2,324.7 tokens per second), 17.3% greater image processing rate, 20.5% faster time to first token (1,006.5ms vs. 1,267.1ms), 30.9% faster time per output token (62.4ms vs. 90.4ms), and 28.0% lower end-to-end latency (3,045.0ms vs. 4,233.8ms).

In the decode-heavy text scenario, the Google G4 VM again demonstrated consistent advantages across both throughput and latency: 9.9% more output token throughput (2,442.7 vs. 2,222.3 tokens per second), 12.5% more input token throughput, 26.8% faster time to first token (77.3ms vs. 105.6ms), 9.8% faster time per output token, and 7.3% lower end-to-end latency.

Results in the prefill-heavy text scenario were mixed. The two platforms produced nearly identical output token throughput (823.5 vs. 822.3 tokens per second) and very similar per-request latency across all metrics. However, the Google G4 VM maintained a 9.4% advantage in input token throughput (39,076.0 vs. 35,699.7 tokens per second), indicating higher capacity for concurrent prefill operations even when per-request latency is comparable.

The architecture behind the results

PT believes the performance differences stem from architectural differences in how the two platforms connect their GPUs in the VM or instance. Google Cloud G4 VMs include dedicated peer-to-peer (P2P) switching infrastructure that enables all eight GPUs to communicate directly at high bandwidth in full all-to-all configuration. AWS G7e instances appear to support high-speed P2P communication only between adjacent GPU pairs; cross-pair inter-GPU traffic must traverse a less direct path, incurring higher latency and/or lower throughput under demanding multi-GPU inference workloads.

This architectural distinction has greater impact in workloads that require sharing intermediary results between GPUs or maximizing the computational power by keeping all eight GPUs running—including multimodal inference with image encoding, large-context prefill processing, and the sustained token generation of decode-heavy sequences.

"For organizations deploying large multimodal LLMs in production—particularly for agentic, image-aware, or long-form generation applications—the choice of cloud infrastructure can have a direct impact on model responsiveness, user experience, and infrastructure cost efficiency," the report states. "In our testing, a Google Cloud G4 VM delivered a significant performance advantage for these workloads over a comparable Amazon EC2 G7e instance."

[Read the full report](#), including detailed methodology and complete results tables.

About Principled Technologies, Inc.

Principled Technologies, Inc. is the leading provider of fact-based marketing and competitive analysis backed by transparent methodologies.

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](#)

This press release can be viewed online at: <https://www.einpresswire.com/article/946686116>

EIN Presswire's priority is source transparency. We do not allow opaque clients, and our editors try to be careful about weeding out false and misleading content. As a user, if you see something we have missed, please do bring it to our attention. Your help is welcome. EIN Presswire, Everyone's Internet News Presswire™, tries to define some of the boundaries that are reasonable in today's world. Please see our Editorial Guidelines for more information.

© 1995-2026 Newsmatics Inc. All Right Reserved.