International Journal of Cloud Applications and Computing
Volume 16 • Issue 1 • January-December 2026

A Comparative Study of Cloud GPU Offerings for AI/ML Engineers

Sanjay P. Ahuja
http://orcid.org/0009-0002-4314-8059
University of North Florida, USA

Madhuri Golanakonda
http://orcid.org/0009-0003-6684-8581
University of North Florida, USA

Sandeep Reddivari
University of North Florida, USA
Received: February 7th, 2026 | **Accepted:** March 23rd, 2026

ABSTRACT

The expanding role of machine learning (ML) and artificial intelligence has become a primary reason behind the demand for high-performance GPUs. Cloud platforms such as Google Cloud Platform, Amazon Web Services, and Microsoft Azure provide scalable access to NVIDIA accelerators. However, variations in GPUs, pricing, usability, and deployment pipelines can be challenging for ML engineers. This paper presents a comprehensive study of GPU platforms across the major cloud providers, covering hardware families (T4, L4, A100, H100, H200, and emerging B200), virtual machine configurations, interconnect technologies, and pricing, including on-demand, spot, and reserved options. Various ecosystem factors were evaluated, including documentation quality, community support, ease of provisioning, and managed ML operations services, such as Vertex AI, SageMaker, and Azure ML. A literature review is provided of MLPerf results and benchmarks to analyze cost, performance, and scalability. The findings provide practical recommendations to guide ML engineers in selecting a suitable cloud for training, inference, and production.

KEYWORDS

Machine Learning, Cloud Computing, GPUs, MLOps, Performance Benchmarking, CI/CD

INTRODUCTION

As artificial intelligence (AI) and machine learning (ML) continue to grow rapidly, the need for powerful computing resources like GPUs has increased. Today's ML tasks, such as image classification, language processing, and generative models, need a lot of computational power to work well. Although on-premises hardware offers raw performance, the high capital cost and maintenance requirements make it impractical for many organizations. Cloud computing has emerged as an alternative that provides scalable access to GPUs on demand without the need for infrastructure.

Google Cloud Platform (GCP), Amazon Web Services (AWS), and Microsoft Azure are the main cloud providers offering GPU-based services. Each one has a range of NVIDIA GPUs, such as T4, L4, A100, H100, and newer models like H200 and B200. For ML engineers, choosing between these

DOI: 10.4018/IJCAC.406737

options can be difficult. GPUs' availability changes, depending on the region, their pricing models vary, and each provider connects its GPUs to different managed services and deployment tools.

ML engineers also need to assess the ease of use of platforms, including the quality of documentation, community support, and set up of resources. Managed ML operations (MLops) services like Vertex AI, SageMaker, and Azure Machine Learning add another layer of complexity by providing built-in tools for training, deployment, and monitoring. As a result, the choice of cloud provider can directly affect models' development, overall costs, and reliability.

Earlier studies have compared GPUs and accelerators across different platforms, and industry groups like MLPerf offer standard means to measure performance. While this research has examined hardware specs and raw performance, there has been less attention to what it is like for developers, how mature the ecosystem is, or how well these tools fit into continuous integration and continuous deployment (CI/CD) workflows. This leaves a gap for ML engineers, who need both technical benchmarks and practical advice on how cloud GPU environments impact everyday work and production systems.

This paper offers a detailed look at GPU options on GCP, AWS, and Azure, with a focus on what matters most to ML engineers. We compared several types of GPU, virtual machine setups, network connections, and pricing. We also examined the support each ecosystem provides, the available managed services, and the deployment process. Using published MLPerf results, we examined performance for cost, inference speed, and scalability. Our results show the trade-offs between performance, cost, and ease of use, and we share practical insights to help ML practitioners choose the best cloud GPU setup for their needs.

BACKGROUND AND RELATED WORK

GPUs for ML

GPUs are now essential for modern ML and deep learning, because they can manage many calculations at once. While CPUs are built for general tasks, GPUs have thousands of cores that make them much better at running matrix multiplications and tensor operations quickly and efficiently.

NVIDIA has been at the forefront of designing GPUs for AI. Volta architecture brought in the first tensor cores, which are built for mixed-precision matrix operations. Turing GPUs, such as the T4, added support for INT8 and INT4 precision, making them great for inference tasks (Reuther et al, 2022). The Ampere A100 improved efficiency with TensorFloat-32, allowing faster training without changing existing models (Yoshida et al, 2022). Hopper GPUs, such as the H100, introduced FP8 precision, which doubles the speed for transformer models like large language models (Luo et al., 2024). The latest H200 and B200 GPUs built on this progress by increasing memory and bandwidth to help with large-scale training.

Tensor cores speed up computations by using lower-precision math without losing accuracy. For example, mixed-precision training uses FP16 or BF16 for calculations but stores the results in FP32 to keep things stable (Reuther et al, 2022). This method can make training two to three times faster than using FP32 alone.

Precision formats are designed to strike a balance between accuracy, computational speed, and memory efficiency. FP32 remains the standard for model training, although it often results in slower processing. FP16 and BF16 formats reduce memory consumption and accelerate model training. TF32, which is supported on Ampere GPUs, provides accuracy comparable to FP32 while achieving performance levels similar to FP16. INT8 and INT4 formats are used primarily during inference to decrease latency and reduce operational costs.

The introduction of the FP8 format further enhances the efficiency of large-scale models. Model training typically requires high computational throughput, substantial memory bandwidth, and support for multiple GPUs such as the A100, H100, or H200. In contrast, inference prioritizes low latency and cost efficiency, making GPUs like the T4 and L4 more suitable. Currently, most training pipelines