North American Journal of Engineering and Research

Est. 2020

Efficiency and Efficacy: AWS Instance Benchmarking of Stable Diffusion 1.4 for AI Image Generation

Author: Yash Jani

Email: yjani204@gmail.com

Abstract

This study benchmarks the performance and cost-efficiency of various AWS instances for AI image generation using the CompVis/stable-diffusion-v1-4 model [1][2]. We evaluate multiple instance types, focusing on performance metrics such as total duration, costs (on-demand, reserved, spot), GPU and memory utilization, temperature, and power draw [3]. Our findings highlight the strengths and weaknesses of each instance type, providing valuable insights for optimizing AI workflows and selecting the most suitable instances. High GPU utilization is emphasized for intensive tasks, while lower temperatures and power draw are noted for sustainability. This analysis empowers researchers, developers, and businesses to maximize AI processing efficiency and manage costs effectively[4].

Introduction

Artificial intelligence (AI) image generation has seen significant advancements with models like CompVis/stable-diffusion-v1-4, which require substantial computational resources to perform efficiently. As AI applications continue to evolve, the demand for scalable and cost-effective computing solutions becomes critical. Amazon Web Services (AWS) offers a diverse range of cloud instances tailored to meet these computational needs, but selecting the appropriate instance type is crucial for optimizing both performance and cost.

This study aims to benchmark various AWS instances specifically for running the CompVis/stable-diffusion-v1-4 model. We focus on key performance metrics, including total duration, costs (on-demand, reserved, and spot), GPU and memory utilization, temperature, and power draw. By analyzing these metrics, we provide insights to help users make informed decisions about the best instance type for their specific AI image generation requirements.

Our evaluation uses standardized benchmarking tools such as MLPerf [5] and NVIDIA's System Management Interface to ensure accurate and detailed performance metrics. This study highlights the strengths and weaknesses of each instance type and offers recommendations for optimizing AI workflows and infrastructure planning. By understanding the trade-offs between performance and cost, users can better manage their AI processing needs in a cost-effective and efficient manner.

Literature Review

- **AI Image Generation Models**

AI image generation has rapidly progressed with the development of advanced models such as Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs). Among these, diffusion models, particularly the CompVis/stable-diffusion-v1-4, have shown remarkable capabilities in generating high-quality images [2]. The stable-diffusion model leverages a series of denoising steps to generate images from noise, offering flexibility and control over the image generation process. Previous studies have demonstrated the effectiveness of stable-diffusion models in various applications, from creative arts to scientific visualization.

- **Benchmarking in AI**

Benchmarking is critical in evaluating the performance of AI models and their deployment environments. The MLPerf [5] benchmarking suite is widely recognized for its standardized tests across different hardware and software configurations, providing a reliable measure of performance. NVIDIA's System Management Interface (nvidia-smi) is another essential tool that offers detailed metrics on GPU utilization, temperature, and power draw [6]. These tools have been extensively used in prior research to benchmark the performance of AI models on various hardware setups.[5]

- **AWS Instances for AI Workloads**

AWS provides a range of instances designed for different computational needs. GPU-optimized instances, such as the P-series and G-series, are particularly suitable for AI workloads due to their high computational power and memory bandwidth. Studies have shown that selecting the right instance type can significantly impact both the performance and cost of AI tasks. For example, the P3 and P4 instances, equipped with NVIDIA V100 and A100 GPUs, respectively, have been highlighted for their superior performance in training deep learning models. However, the high costs

associated with these instances necessitate a careful cost-performance analysis.

- **Performance Metrics and Cost Analysis**

Evaluating the performance of AWS instances involves multiple metrics, including total duration, GPU and memory utilization, and power efficiency. Previous research has emphasized the importance of these metrics in understanding the trade-offs between performance and operational costs [7]. High GPU utilization is often associated with better performance but can lead to increased power consumption and heat generation, impacting sustainability and operational costs [6]. Studies have also explored the benefits of reserved and spot pricing models in reducing overall expenses for long-term AI projects.

- **Contributions of This Study**

While there is extensive literature on AI image generation and benchmarking, specific studies on the performance and cost-efficiency of AWS instances using the stable-diffusion-v1-4 model are limited. This study fills this gap by providing a comprehensive benchmarking analysis, focusing on both performance and sustainability metrics. Our findings offer valuable insights for researchers, developers, and businesses aiming to optimize their AI infrastructure for efficient and cost-effective image generation.

Methodology

Overview

This study benchmarks the performance and cost-efficiency of various AWS instances for AI image generation using the CompVis/stable-diffusion-v1-4 [1] model. Our evaluation focuses on several key performance metrics, including total duration, costs (on-demand, reserved, and spot), GPU and memory utilization, temperature, and power draw. We employed MLPerf [5] and NVIDIA's System Management Interface (nvidia-smi) to ensure detailed and accurate performance metrics for each instance type.

AWS Instances [8]

We selected a range of GPU-optimized AWS instances, each with different computational capacities to provide a comprehensive comparison:

- **P5.48xlarge:**
  - **vCPUs**: 192
  - **Memory**: 2048 GiB
  - **GPU**: 8 NVIDIA H100
- **P4d.24xlarge:**
  - **vCPUs**: 96
  - **Memory**: 1152 GiB
  - **GPU**: 8 NVIDIA A100
- **P3.2xlarge:**
  - **vCPUs**: 8
  - **Memory**: 61 GiB
  - **GPU**: 1 NVIDIA V100 Tensor Core
- **G6.xlarge:**
  - **vCPUs**: 4
  - **Memory**: 16 GiB
  - **GPU**: 1 NVIDIA L4
- **G5.xlarge:**
  - **vCPUs**: 4
  - **Memory**: 16 GiB
  - **GPU**: 1 NVIDIA A10G
- **G4dn.large:**
  - **vCPUs**: 4
  - **Memory**: 16 GiB
  - **GPU**: 1 NVIDIA T4 Tensor Core
- **G3s.xlarge:**
  - **vCPUs**: 4
  - **Memory**: 30.5 GiB
  - **GPU**: 1 NVIDIA Tesla M60

Data Collection

Data was collected by running the CompVis/stable-diffusion-v1-4 [1] model on each instance located in the Ohio region to generate 30 images with a size of 512*512. The following metrics were recorded:

Performance Metrics

- **Total Duration:**
  - Measured in minutes, each instance's time required to complete the image generation task.
- **Costs:**
  - Costs were evaluated for different pricing models (on-demand, reserved, and spot) in the Ohio region.
- **GPU and Memory Utilization:**
  - Utilization rates were recorded to assess how effectively each instance utilized its resources.
- **Temperature and Power Draw:**
  - Average GPU temperature and power draw were monitored to evaluate each instance's energy efficiency and cooling requirements.

Benchmarking Tools

- **MLPerf [5]:**
  - Used to ensure standardized benchmarking across different hardware configurations.
- **NVIDIA System Management Interface (nvidia-smi):**

- Provided detailed metrics on GPU utilization, temperature, and power draw, ensuring accurate and comprehensive data collection.

Data Analysis

The collected data was analyzed to compare the performance, cost efficiency, and energy efficiency of each AWS instance type. Bar charts and heatmaps were used to visualize total duration, costs, GPU utilization, and power draw. Correlation matrices were created to understand the relationships between different performance metrics.

Implications

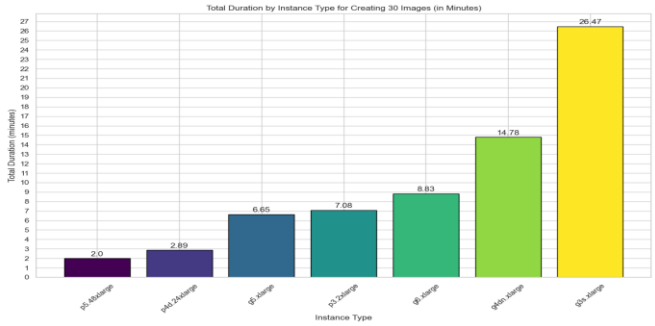
This methodology allows us to provide detailed insights into the strengths and weaknesses of each AWS instance type, helping users optimize their AI workflows and select the best instance for their specific needs and budget

Result

Total Duration (in Minutes)

The bar chart visualizes the total duration required by different AWS instance types to generate 30 images using the CompVis/stable-diffusion-v1-4 [1] model. The durations are measured in minutes, clearly comparing each instance type's performance.

- **P5.48xlarge:** This instance, equipped with NVIDIA H100 GPUs, delivers the fastest processing time, making it ideal for high-demand AI tasks.



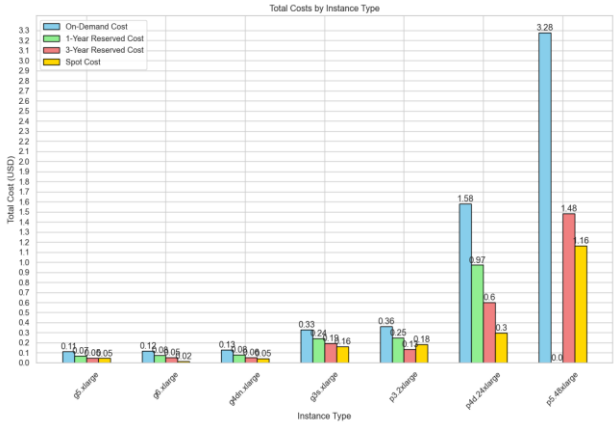
- **P4d.24xlarge:** With 8 NVIDIA A100 GPUs, this instance offers excellent performance. It is slightly slower than the P5.48xlarge but still highly efficient.
- **P3.2xlarge:** Featuring a single NVIDIA V100 GPU, this instance provides moderate performance suitable for less intensive AI tasks.
- **G6.xlarge:** This instance is slower than the G5.xlarge, highlighting its reduced efficiency, and it is equipped with an NVIDIA L4 GPU.
- **G5.xlarge:** This instance, with an NVIDIA A10G GPU, offers a good balance between performance and cost, performing slightly faster than the P3.2xlarge.
- **G4dn.xlarge:** This instance, using an NVIDIA T4 Tensor Core GPU, shows considerably slower performance compared to the G6.xlarge.
- **G3s.xlarge:** This comparison's slowest instance indicates the limited capability of the older NVIDIA Tesla M60 GPU.

Analysis

The performance differences among the AWS instances underscore the significant impact of GPU capabilities on AI image generation tasks. Interestingly, the G5.xlarge instance outperforms both the G6.xlarge and P3.2xlarge instances, despite expectations based on GPU specifications. This indicates that the NVIDIA A10G GPU in the G5.xlarge may be more efficient for this particular workload than the NVIDIA L4 GPU in the G6.xlarge and the NVIDIA V100 GPU in the P3.2xlarge.Instances with more powerful GPUs, such as the NVIDIA H100 in the P5.48xlarge, dramatically reduce processing times, offering high efficiency for demanding workloads. Conversely, instances with older or less powerful GPUs, like the NVIDIA Tesla M60 in the G3s.xlarge, experience much longer durations. This analysis highlights the importance of selecting the appropriate GPU to optimize both performance and cost-efficiency for specific AI tasks. Additionally, instances with fluctuating utilization rates could benefit from optimization and concurrent usage strategies to maximize efficiency.

Total Costs by Instance Type

The bar chart displays the total costs associated with different AWS instance types for running the CompVis/stable-diffusion-v1-4 [1] model to generate 30 images. The costs are divided into four categories: On-Demand, 1-Year Reserved, 3-Year Reserved, and Spot.

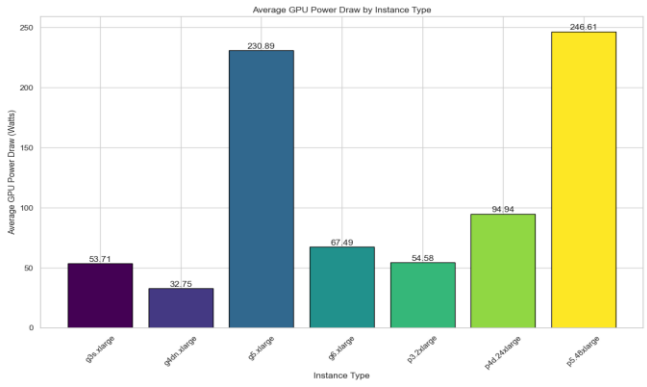


- **P5.48xlarge:** The highest costs among all instances reflect its top-tier performance with NVIDIA H100 GPUs. Suitable for projects where performance is critical and budget is less of a concern.
- **P4d.24xlarge:** This instance is much more expensive, justified by its superior performance with 8 NVIDIA A100 GPUs. It is best used for high-demand, long-term projects where performance is critical.
- **P3.2xlarge:** The P3.2xlarge instance has higher costs across all pricing models due to its superior performance. It is more suitable for specific tasks that require its capabilities but less cost-efficient for general use.
- **G6.xlarge:** While slightly more expensive on-demand than G5.xlarge, G6.xlarge offers the lowest spot cost, making it a very economical choice for spot usage. However, its overall performance is lower than G5.xlarge.
- **G5.xlarge:** This instance provides an excellent balance between performance and cost, particularly in spot pricing. It is cheaper in on-demand and reserved pricing compared to other instances with similar or lower performance.
- **G4dn.xlarge:** The G4dn.large instance has higher on-demand costs compared to G5.xlarge and G6.xlarge but offers competitive reserved and spot pricing. Its performance is lower, making it less attractive for high-performance needs.
- **G3s.xlarge:** This instance has significantly higher on-demand and reserved costs, reflecting its older and less efficient GPU. Despite competitive spot pricing, it is less cost-effective for regular use.

Analysis

The cost comparison across different instance types highlights the trade-offs between performance and cost efficiency. Instances like G5.xlarge and G6.xlarge offer balanced and economical options, while high-performance instances such as P4d.24xlarge and P5.48xlarge provide substantial savings through reserved and spot pricing, making them suitable for intensive and long-term projects.

Average GPU Power Draw



The bar chart illustrates the average GPU power draw in watts for different AWS instance types during image generation using the CompVis/stable-diffusion-v1-4 [1] model. Understanding the power draw is crucial for evaluating the energy efficiency of each instance type, which can have implications for operational costs and environmental impact.

- **P5.48xlarge:** The highest power draw among all instances, which is consistent with its top-tier performance. This instance requires substantial energy infrastructure and cooling management.
- **P4d.24xlarge:** Moderate power draw, making it more energy-efficient than the P5.48xlarge and G5.xlarge, while still offering excellent performance.
- **P3.2xlarge:** Lower power draw, indicating good energy efficiency. Suitable for tasks that do not require the highest performance levels.
- **G6.xlarge:** Moderate power draw, slightly higher than G3s.xlarge but still energy-efficient, making it a good balance between performance and power consumption.
- **G5.xlarge:** This instance shows a high power draw, reflecting its strong performance capabilities. However, it also indicates higher energy costs and the need for robust cooling solutions.
- **G4dn.xlarge:** The lowest power draw among all instances, making it extremely cost-effective and environmentally friendly for long-term operations.
- **G3s.xlarge:** One of the lowest power draws, indicating high energy efficiency. Suitable for light workloads and long-term use with lower operational costs.

Analysis

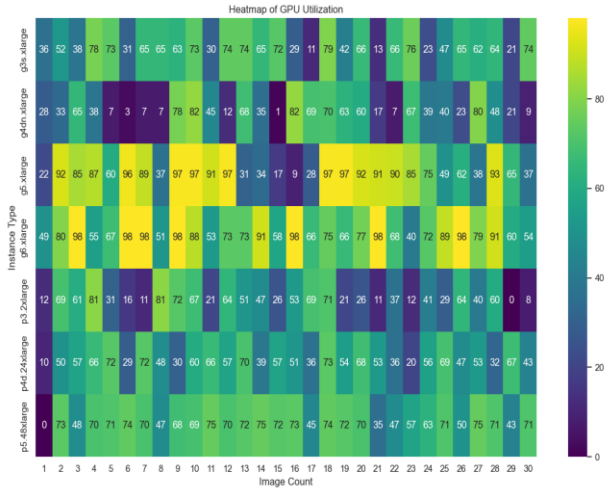
Understanding the average GPU power draw of AWS instances is crucial for:

- **Cost Efficiency:** Lower power consumption reduces energy costs, making instances like G4dn.xlarge and G3s.xlarge cost-effective for long-term use.
- **Environmental Impact:** Lower energy consumption supports sustainability and reduces carbon footprints.
- **Operational Planning:** High-power draw instances like P5.48xlarge and G5.xlarge require robust cooling and energy infrastructure, which is important for data center management.
- **Performance vs. Efficiency:** High-performance instances consume more power, so balancing performance and energy efficiency is essential when selecting instances for AI workloads.

Heatmap of GPU Utilization

The heatmap shows GPU utilization percentages for different AWS instances as image counts increase. Darker colors indicate lower utilization and lighter colors indicate higher utilization.

- **P5.48xlarge:** Utilization varies from 0% to 75%, indicating potential inefficiencies. Low utilization periods can support multiple AI pipelines simultaneously.
- **P4d.24xlarge:** High utilization (50%-60%) from image counts 8 to 20, with some lower periods. Further optimization could enhance consistent high utilization.
- **P3.2xlarge:** Utilization ranges from 11% to 81%, with dips around counts 6 and 12. Optimization could improve handling low periods, enabling multiple AI models to run simultaneously.



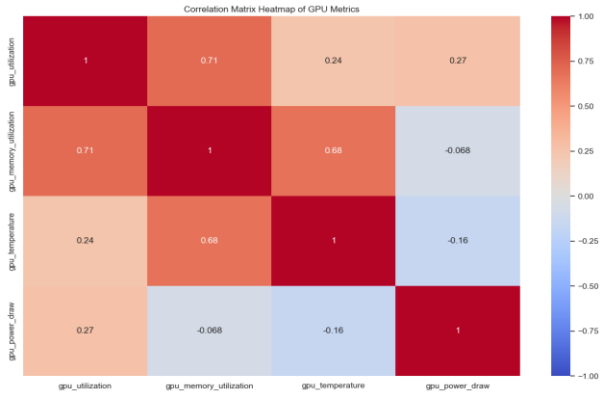
- **G6.xlarge:** Shows high utilization (mostly above 70%) from image counts 5 to 20, with few fluctuations. It effectively uses GPU resources, and low periods could support extra AI models.
- **G5.xlarge:** Maintains high utilization (90%-97%) between image counts 10 and 20, showing strong performance. Minor dips suggest opportunities for concurrent pipeline usage.
- **G4dn.xlarge:** Utilization varies from 3% to 85%, with significant drops. Optimization could improve performance, and low periods could support additional AI models.
- **G3s.xlarge:** Utilization fluctuates from 29% to 79%, indicating inconsistent performance and potential inefficiencies. It may benefit from optimization or running multiple AI pipelines during low utilization.

5. Correlation Matrix Heatmap of GPU Metrics [9]

The correlation matrix heatmap illustrates relationships between GPU metrics for various AWS instances, with coefficients ranging from -1 to 1.

Analysis:

- **GPU Utilization & Memory Utilization:** Strong positive correlation (0.71) - higher GPU activity increases memory usage.
- **GPU Utilization & Temperature:** Weak positive correlation (0.24) - slight temperature increase with higher GPU use.
- **GPU Utilization & Power Draw:** Weak positive correlation (0.27) - power consumption rises with GPU use.



- **Memory Utilization & Temperature:** Moderate positive correlation (0.68) - more memory usage raises temperatures.
- **Memory Utilization & Power Draw:** Very weak negative correlation (-0.068) - almost no link between memory use and power consumption.
- **Temperature & Power Draw:** There is a weak negative correlation (-0.16)—higher temperatures don't strongly correlate with higher power consumption.

Implications:

- **Resource Utilization:** Optimize by running multiple pipelines during low GPU use periods.
- **Performance Monitoring:** Identify bottlenecks and improve GPU usage.
- **Thermal Management:** Enhance cooling for instances with high utilization-temperature correlation to prevent throttling.
- **Energy Efficiency:** Optimize power draw to reduce operational costs.

Conclusion:

This study benchmarks various AWS instances for AI image generation using the CompVis/stable-diffusion-v1-4 [1] model, providing insights into performance, cost, and

energy efficiency. Our analysis highlights the following key findings:

- **P5.48xlarge:**
  - Equipped with NVIDIA H100 GPUs, this instance delivers the fastest processing times, making it ideal for high-demand AI tasks. However, its high cost makes it suitable for projects where speed is critical and budget is less of a concern.
- **P4d.24xlarge:**
  - With 8 NVIDIA A100 GPUs, this instance offers excellent performance, slightly slower than the P5.48xlarge. Reserved and spot pricing provide significant cost savings, making it economical for long-term projects.
- **P3.2xlarge:**
  - Featuring a single NVIDIA V100 GPU, this instance provides moderate performance suitable for less intensive AI tasks. Its affordability makes it ideal for smaller-scale projects.
- **G6.xlarge:**
  - The reasonable performance provided by the NVIDIA L4 GPU makes this instance suitable for general AI tasks and an economical choice for moderate AI workloads.
- **G5.xlarge:**
  - This instance offers balanced performance with an NVIDIA A10G GPU, delivering decent speed at a lower cost. It is particularly cost-effective in reserved and spot pricing models, making it ideal for budget-conscious users.
- **G4dn.large:**
  - Equipped with an NVIDIA T4 Tensor Core GPU, this instance performs light to moderate AI tasks satisfactorily. Its energy efficiency and cost-effectiveness stand out, making it ideal for cost-sensitive projects.
- **G3s.xlarge:**
  - This instance features an older NVIDIA Tesla M60 GPU, which provides the least performance but is the most affordable. It is best suited for light workloads or experimental projects.

Overall, the P5.48xlarge instance excels in speed but at a high cost, making it suitable for critical high-performance tasks. The G5.xlarge balances cost and performance, making it a versatile choice for many users. Instances like the G4dn.large stand out for their energy efficiency, making them ideal for long-term cost-effective use. These insights help researchers, developers, and businesses optimize their AI infrastructure by balancing performance, cost, and energy efficiency according to their specific needs.

By providing detailed benchmarking data, this study empowers users to make informed decisions about selecting an AWS instance for AI image generation, ultimately enhancing efficiency and cost management in AI workflows.

Future work

This study provides valuable insights into the performance and cost-efficiency of various AWS instances for AI image generation using the CompVis/stable-diffusion-v1-4 model. However, several avenues for future research could further enhance our understanding and optimize AI workflows:

- **Extended Benchmarking Across More Models:**

Benchmark additional AI models, such as other diffusion models, GANs, and VAEs, to compare performance across various types of AI image generation tasks.

- **Real-World Application Testing:**

Implement benchmarks in real-world applications, such as automated content creation, video game design, and medical imaging, to validate findings with practical use cases.

- **Longitudinal Cost, Energy Efficiency and Sustainability Analysis:**

Conduct long-term cost analysis to understand implications of fluctuating spot prices and evolving AWS offerings. Investigate AI training's environmental impact by incorporating renewable energy and optimizing for lower carbon footprints. Use energy-efficient hardware and software for sustainable practices.

References

[1] CompVis, “CompVis”, Jan. 4, 2023. [Online]. Available: https://huggingface.co/CompVis [Accessed: Jun. 8, 2024]

[2] Joe-Wong, X Z W J. (2020, July 1). Machine Learning on Volatile Instances. https://arxiv.org/pdf/2003.05649v1.pdf

[3] Alex, D P N. (2021, May 11). Diffusion Models Beat GANs on Image Synthesis. https://arxiv.org/abs/2105.05233v4

[4] Zhang, H., Li, Y., Xiao, W., Huang, Y., Di, X., Yin, J., See, S., Luo, Y., Lau, C T., & You, Y. (2023, January 1). MIGPerf: A Comprehensive Benchmark for Deep Learning Training and Inference Workloads on Multi-Instance GPUs. https://doi.org/10.48550/arxiv.2301.00407

[5] MLPerf. MLPerf. https://mlperf.org, 2019. [Accessed: Jun. 8, 2024].

[6] Santhi, Y A E E A E A B C E. (2020, May 11). Verified instruction-level energy consumption measurement for NVIDIA GPUs. https://arxiv.org/pdf/2002.07795v1.pdf

[7] “From Words to Watts: Benchmarking the Energy Costs of Large Language Model Inference” https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=10363447

[8] Amazon EC2 Instances, "Amazon EC2 Instances," Jan. 1, 2023. [Online]. Available: https://aws.amazon.com/ec2/.

[9] E. K. S. U-Ruekolan, "Efficient large Pearson correlation matrix computing using hybrid MPI/CUDA," n.d. [Online]. Available: https://ieeexplore.ieee.org/document/5930127/. [Accessed: Jun. 8, 2024].