DELL Technologies

Whitepaper

Generative AI and Stable Diffusion Image Generation on the Dell PowerEdge XE9680 Server

Authors:
Ben Fauber, Ph.D.,
Senior AI Research Scientist
and Distinguished Technical Staff

Bhavesh Patel,
Senior Distinguished Engineer

Introduction

Generative artificial intelligence (AI) is impacting many aspects of the business community. ChatGPT and other similar large language models (LLMs) have captured attention for their amazing ability to create human-like prose.[1,2] Additionally, generative AI can create visually captivating artistic content, encompassing images, videos, and audio. Langevin diffusion deep learning models are primarily employed to generate this content, with open-source image generation models, such as Stable Diffusion, being the most popular approach.[3]

Background

Computer vision (CV) research has focused on image recognition algorithms for image classification[4] and image segmentation[5] to gain a better understanding of the objects present within them. Image segmentation has applications most closely aligned to autonomous vehicles and other real-time image processing applications such as digital manufacturing, healthcare analysis, physical security, and sports analytics. Additionally, there has been a significant focus on integrating computer vision with natural language processing (NLP) algorithms to generate captions/summaries for images based on their content.[6]

Convolutional neural networks (CNNs) were employed in earlier computer vision models to establish the relationships between images and the model objectives of image classification and segmentation. Subsequently, computer vision models have advanced to leverage both U-net architectures based on convolutional neural networks[7] and attention-based transformers.[8]

Generating Images from Text Prompts

A significant advancement in computer vision research occurred with the development of contrastive learning image pairs (CLIP). CLIP leverages text annotations and captions for images to train deep learning (DL) neural networks (NN) and generative AI networks on millions to billions of image/text pairs.[9] The resulting output of the training process is a neural network that models the associations between text and images. As a result, these generative AI CLIP models can create rich images and visual representations from a single line of text.

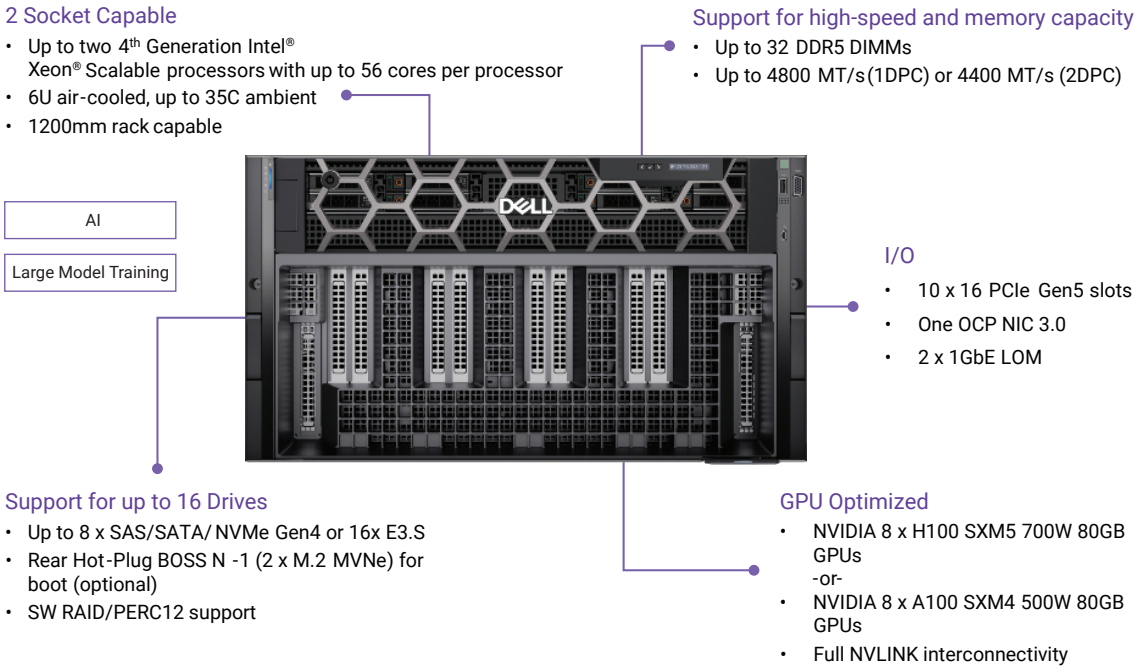
DALL-E from OpenAI was one of the early breakthroughs in the field of CLIP models,[10] followed by DALL-E2.[11] In mid-2022, Stability AI collaborated with an academic institution and financed the development of an open-source generative AI image model, like DALL-E2, resulting in the Stable Diffusion model.[3] Following the release of the Stable Diffusion model, enterprises, academics, and individuals have contributed to the development of this broad class of text-to-image and image-to-image generative AI image generation tools.[12]

Dell Technologies PowerEdge XE9680 Server

The Dell PowerEdge XE9680 is a high-performance server designed and optimized to enable uncompromising performance for artificial intelligence, machine learning, and high-performance computing workloads. Dell PowerEdge is launching our innovative 8-way GPU platform with advanced features and capabilities. We have previously demonstrated the PowerEdge XE9680 server's impressive capabilities when launching and inferencing large language models of comparable scale to ChatGPT.[13]

- 8x NVIDIA® H100 80GB 700W SXM GPUs or 8x NVIDIA® A100 80GB 500W SXM GPUs
- 2x Fourth Generation Intel® Xeon® Scalable Processors
- 10x PCIe Gen 5 x16 FH Slots
- 8x SAS/NVMe SSD Slots (U.2) and BOSS-N1 with NVMe RAID

Dell PowerEdge XE9680



	PowerEdge XE9680
CPU	2x Intel® Xeon® 8470 52-core Processor
GPU	8x NVIDIA® H100-SXM-80GB (700W)
System Memory	32x64GB – 2TB
Host NIC	NVIDIA® CX7

The PowerEdge XE9680 6U server is designed for AI, machine learning, and deep learning applications. It features the latest Intel® Xeon® processors with up to 56 cores, 8 NVIDIA® H100 or A100 GPUs, NVIDIA® NVLink™ technology for GPU-GPU communication, and supports up to 4 TB RDIMM of CPU RAM. The server supports virtualization options like NVIDIA® Multi-Instance GPU (MIG) capability, DDR5, NVLink™, PCIe Gen 5.0, and NVMe SSDs. It also supports NVIDIA® GDS (GPUDirect Storage), which provides a direct data path between GPU memory and storage, increasing system bandwidth and decreasing latency. The server is certified by NVIDIA® and has a secure, efficient, and comprehensive systems management solution with the OpenManage Enterprise console and iDRAC.

The results described in this article used a XE9680 server with 8 x 80 GB H100 NVIDIA® HGX GPU cards with NVLink™ technology, 2 TB of CPU RAM, and 2 x Intel® Xeon® processors on each server. The XE9680 server was configured with the Ubuntu v22.04 Linux operating system, Anaconda v23.1.0, CUDA v12.1, cuDNN v8.8.1, and the python dependencies required for Stable Diffusion and the HuggingFace library to enable Stable Diffusion, known as the Diffusers library.[12] A full list of the python dependencies installed on the XE9680 server can be found in the Appendix section of this article.

Image Generation on the PowerEdge XE9680 Server

Dell Technologies has demonstrated that state-of-the-art (SOTA) image generation models can be launched and queried on our PowerEdge XE9680 server platform. Image generation models require sizable memory and GPU capabilities at inference. In this instance, the open-source image generation models included variations of Stable Diffusion,[3] via the HuggingFace Diffusers library,[12] were launched, queried, and benchmarked on a PowerEdge XE9680 server.

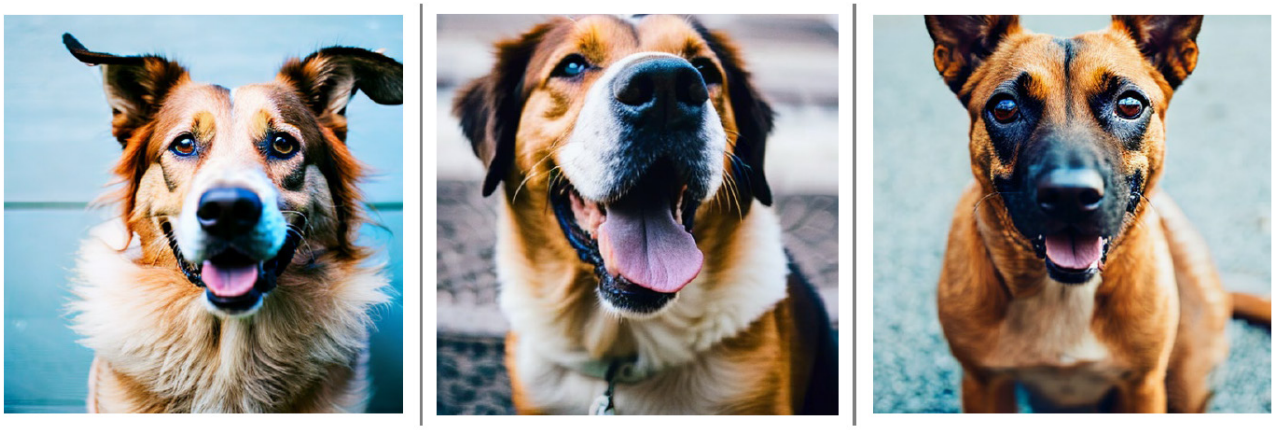


Figure 1. Images generated with text prompt = "Portrait of happy dog, close up," using the HuggingFace Diffusers text-to-image model with batch size = 1, number of iterations = 25, float16 precision, DPM Solver Multistep Scheduler, and Stable Diffusion v2.1.

Image generation via a text prompt, also known as text-to-image (text2image) was initiated with the HuggingFace Diffusers library for single image generation. Example model outputs are shown in Figure 1, where the text prompt = "Portrait of happy dog, close up" was provided to the Stable Diffusion v2.1 image generation model. The 512 x 512 resolution images in Figure 1 were generated in 0.64 seconds per image by the model running on an XE9680 server.

An assessment of the image generation latency for the Diffusers text-to-image model on the PowerEdge XE9680 server, as the image resolution was varied, is shown in Figure 2. Images of *n* x *n* resolution, up to 1,024 x 1,024 resolution, were generated by the XE9680 in seconds or less. Very high-resolution images, such as 2,048 x 2,048 resolution, were generated on the XE9680 server in just 16 seconds. Attempts to generate even higher-resolution images, such as 4,096 x 4,096 resolution, resulted in out-of-memory (OOM) errors due to the memory requirements of performing such a large image generation calculation on a single GPU. At the time this article was written, multi-GPU processing for a single workload was not available for single or batch image generation.

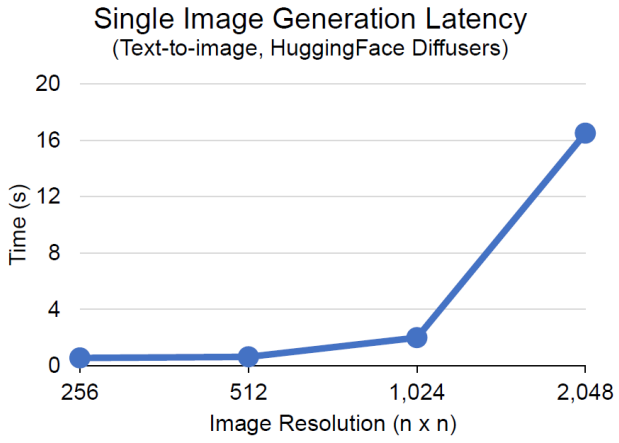


Figure 2. Image generation algorithm runtimes versus n x n image resolution on the Dell PowerEdge XE9680 server. Larger images require more time for image generation. All images used the same HuggingFace text-to-image Diffusers text prompt = "Portrait of happy dog, close up," batch size = 1, number of iterations = 25, float16 precision, DPM Solver Multistep Scheduler, and Stable Diffusion v2.1.

Variations of Image Generation on the PowerEdge XE9680 Server

Image generation with diffusion techniques have advanced to the point where they can transform a text prompt into new images, alter the composition of an existing image to create a new one (Figure 3), modify features within an image (e.g., foreground, background, etc.) (Figure 4), add new elements to an existing image, create panoramic images (Figure 5), and generate new video content.[12] These methods are also capable of increasing the size of an existing image to 2-4 times its original scale.

These image generation and alteration methods can be applied with the Dell PowerEdge XE9680 server and the HuggingFace Diffusers library to generate new images in just one second. Generating multiple images and merging them, via inpainting processes, to create a new high-resolution panoramic image from a text prompt does require multiple seconds of algorithm runtime (Figure 5).

Figure 3. Style transfer applied to input image (left), at 512 x 768 resolution, and text prompt = "Realistic landscape", to generate a new image (right), at 512 x 768 resolution, using Stable Diffusion v1.5 with HuggingFace Diffusers Img2Img. Batch size = 1, number of iterations = 50, float16 precision, and algorithm runtime = 1.1 seconds.

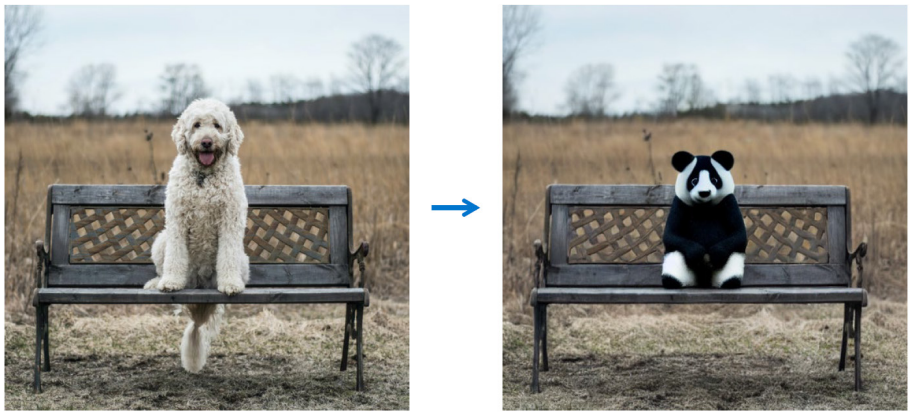


Figure 4. Inpainting technique applied to input image (left), at 512 x 512 resolution, and text prompt = "Panda bear, sitting on park bench", to generate new image (right), at 512 x 512 resolution, using Stable Diffusion v1.5 with HuggingFace Diffusers Inpaint. Batch size = 1, number of iterations = 50, float16 precision, and algorithm runtime = 1.1 second.

Figure 5. New 512 x 2,048 resolution panoramic image generated with text prompt = "Dolomite mountains," and HuggingFace Diffusers Panorama with Stable Diffusion v2. Batch size = 1, number of iterations = 50, float16 precision, DDIM Scheduler, and algorithm runtime = 27 seconds.

Batch Image Generation on the PowerEdge XE9680 Server

To generate new images during inference, diffusion methods typically rely on large clusters of synchronized graphics processing units (GPUs) and extensive hardware runtime. However, by utilizing the Dell PowerEdge XE9680 server, these generative AI processes can be significantly accelerated, enabling the creation of tens to hundreds of pictures at reasonable resolutions in mere seconds (Figure 6). Moreover, large high-resolution images (e.g., 2,096 x 2,096 pixels) can be generated from a single text prompt in just a few seconds by harnessing the power of the PowerEdge XE9680 server.

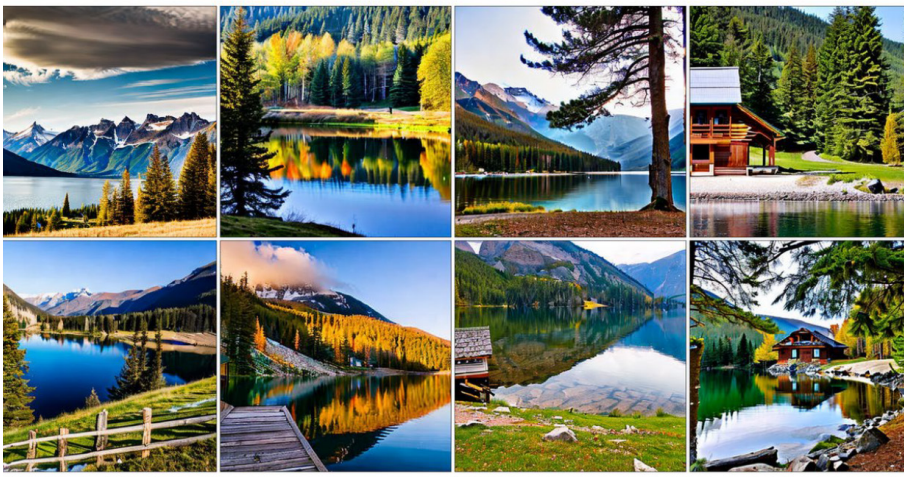


Figure 6. A batch of eight images generated with text prompt = "Photo of cabin on a lake near mountains," using HuggingFace Diffusers text-to-image with batch size = 8, number of iterations = 25, float16 precision, DPM Solver Multistep Scheduler, and Stable Diffusion v2.1.

An assessment of the batch image generation latency for the HuggingFace Diffusers text-to-image model on the PowerEdge XE9680 server, as both the batch size and image resolution were varied, are shown in Figure 7. Images up to 2,048 x 2,048 resolution were generated by the XE9680 server. Larger batch sizes increased the memory requirements on the infrastructure, thus exceptionally large batch sizes could be achieved with smaller resolution images (i.e., 256 x 256). Whereas very high-resolution images (i.e., 2,048 x 2,048) could only generate a batch size of one due to memory constraints.

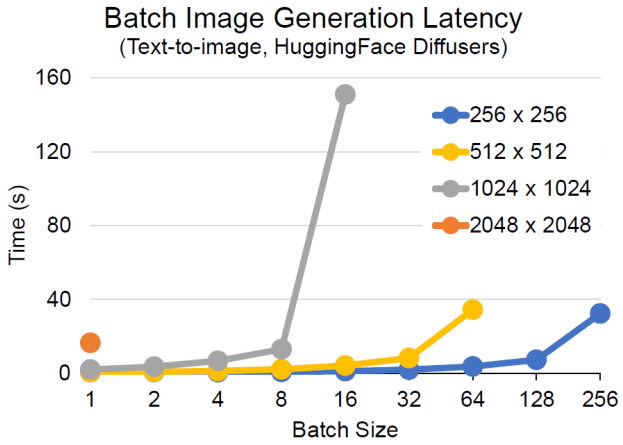


Figure 7. Batch image generation times for images of n x n resolution on the Dell PowerEdge XE9680 server. All images used the same HuggingFace text-to-image Diffusers text prompt = "Photo of cabin on a lake near mountains," number of iterations = 25, float16 precision, DPM Solver Multistep Scheduler, and Stable Diffusion v2.1.

Focusing on the batch image generation of images at both 256 x 256 and 512 x 512 resolutions shows the powerful capabilities of the Dell PowerEdge XE9680 server to rapidly generate multiple images within seconds (Figure 8). This capability accelerates the evaluation, prompt tuning, and further evaluation required for creative design cycles. These results demonstrate how batches of 32 images at 512 x 512 resolution can be generated with the HuggingFace Diffusers package on the Dell PowerEdge XE9680 server in less than 10 seconds. Further, batches of 64 images at 256 x 256 resolution can be created in less than 5 seconds with the same platform, enabling rapid prototyping and creative design cycles for businesses and professionals.

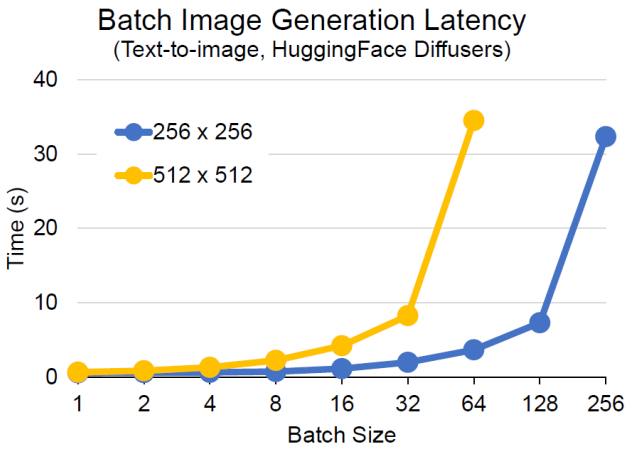


Figure 8. Batch image generation times for images of both 256 x 256 (blue) and 512 x 512 (gold) resolutions on the Dell PowerEdge XE9680 server. All images used the same HuggingFace text-to-image Diffusers text prompt = "Photo of cabin on a lake near mountains," number of iterations = 25, float16 precision, DPM Solver Multistep Scheduler, and Stable Diffusion v2.1.

Benchmarking Image Generation on the PowerEdge XE9680 Server

The PowerEdge XE9680 server was benchmarked against image generation latency values provided by Lambda.[14] For this benchmarking exercise, a single NVIDIA® H100 GPU with 80 GB RAM on the Dell PowerEdge XE9680 server was compared against the NVIDIA® H100 GPU Stable Diffusion benchmark published by Lambda's ML Labs team. The benchmark code provided by Lambda Labs, as well as the same dependencies and HuggingFace Diffusers pipeline instantiation was run on the PowerEdge XE9680 server.

In this study, it was noted at the XE9680 outperformed and with approximately 2-fold greater throughput than the Lambda benchmarks (Figure 9). The results of our benchmark study are presented in the same format of images per second data reporting provided by Lambda, and the data in Figure 9 is directly from the Lambda NVIDIA® H100 GPU study under identical software-defined image generation conditions – only the hardware is different. It was noted that the Lambda server contained a PCIe form factor NVIDIA® H100 GPU, whereas the Dell PowerEdge XE9680 server contained a HGX form factor NVIDIA® H100 GPU. This difference in form factors might account for some of the performance difference between the Dell and Lambda hardware.

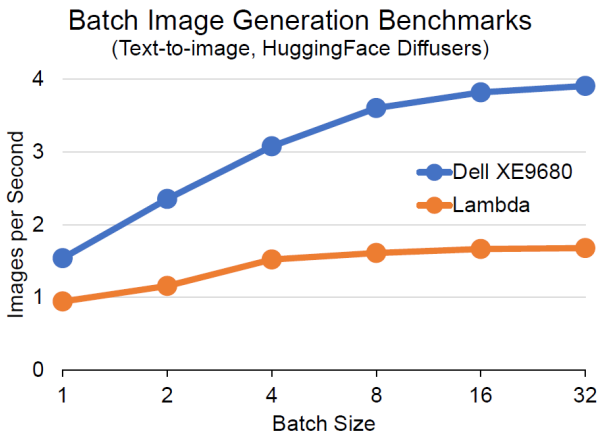


Figure 9. Batch image generation latency values for both Dell PowerEdge XE9680 server (blue) and Lambda server (orange). Both studies used a single NVIDIA® H100 GPU with 80 GB GPU RAM, where the GPU form factor for the Dell server GPU was HGX and the Lambda server GPU was PCIe. All images used the same HuggingFace text-to-image Diffusers PyTorch code with text prompt = "a photo of an astronaut riding a horse on mars," number of iterations = 30, 512 x 512 image resolution, float16 precision, DDIM Scheduler, Stable Diffusion v1.4.

Business Impact of Image Generation Models

Stable Diffusion image generation has the potential to significantly impact businesses across multiple industries. By enabling creatives to rapidly prototype and fine-tune campaigns, this technology can help businesses reduce their time-to-market and improve the effectiveness of their marketing and advertising efforts. Early adopters of this technology have included professionals in advertising, marketing, ideation, film, special effects, photography, and art.

Architecture firms are fully embracing the potential of image generation models. Zaha Hadid Architects (ZHA) recently reported that most of their projects are using image generation models for ideation and project design.[15] Patrick Schumacher, a ZHA studio principal, noted that, "…you can generate ideas with clients and within the team, because of light, shadow, geometry, coherency, the sense of gravity and order is so potent, and the ideas are still striking." Further, ZHA estimated that 10-15% of the generated images were being carried forward into the next stage of design which involves 3D modeling.

In industries such as film, special effects, and photography, image generation can help reduce production costs and accelerate the time required to develop visual effects and renderings. This can enable filmmakers and photographers to bring their creative visions to life more efficiently and with greater accuracy.

The interior design and real estate industries can use image generation to create virtual staging and visualizations of properties. This can help potential buyers or renters visualize what a space could look like, even before any physical changes are made. This can ultimately help real estate professionals to close deals more quickly and effectively.

Conclusion

Stable Diffusion image generation is a powerful technology that has the potential to transform the way businesses approach marketing, advertising, film, special effects, photography, art, interior design, real estate, and other creative pursuits. Its ability to rapidly prototype and fine-tune campaigns, create virtual staging and visualizations, and reduce production costs can lead to improved efficiency, creativity, and cost-effectiveness. As technology continues to evolve and improve, we can expect to see it applied to even more industries and use cases, driving innovation and growth for businesses across industries and verticals.

REFERENCES

1. Fauber, B. "Unleashing the power of large language models like ChatGPT for your business." 2023, Dell Technologies white paper, https://www.delltechnologies.com/asset/en-us/solutions/infrastructure-solutions/industry-market/unleashing-the-power-of-large-language-models-fauber.pdf (accessed 11Apr2023).
2. Open AI Research. "ChatGPT." 2022, https://chat.openai.com/
3. Rombach, et al. "High-resolution image synthesis with latent diffusion models." 2022, arxiv.org/abs/2112.10752.
4. Peng, et al. "A survey: Image classification models based on convolutional neural networks." IEEE International Conference on Computer Research and Development 2022, 291-298.
5. Minaee, et al. "Image segmentation using deep learning: A survey." 2000, arxiv.org/abs/2001.05566.
6. Venugopalan, et al. "Captioning images with diverse objects." 2016, arxiv.org/abs/1606.07770.
7. Shelhamer, et al. "Fully convolutional networks for semantic segmentation." IEEE Transactions on Pattern Analysis and Machine Intelligence 2017, 39(4), 640 - 651.
8. Dosovitskiy, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." International Conference on Learning Representations Conference (ICLR) 2021.
9. Radford, et al. "Learning transferable visual models from natural language supervision." International Conference on Machine Learning (ICML) 2021.
10. Ramesh, et al. "Zero-shot text-to-image generation." International Conference on Machine Learning (ICML) 2021.
11. Ramesh, et al. "Hierarchical text-conditional image generation with CLIP latents." 2022, arxiv.org/abs/2204.06125.
12. HuggingFace Inc. "Diffusers." https://huggingface.co/docs/diffusers/index (accessed 15Apr2023).
13. Fauber and Patel. "Launching & running large language models on a single Dell server produces outstanding results." 2023, Dell Technologies white paper, https://www.delltechnologies.com/asset/en-us/products/servers/industry-market/launching-llms-on-poweredge-xe9680.pdf (accessed 03May2023).
14. Pinkney, J. "Lambda Diffusers." 2023, Lambda Labs, https://github.com/LambdaLabsML/lambda-diffusers (accessed 03May2023).
15. Barker, N. "ZHA developing "most" projects using AI-generated images says Patrik Schumacher." 2023, dezeen Magazine, https://www.dezeen.com/2023/04/26/zaha-hadid-architects-patrik-schumacher-ai-dalle-midjourney/ (accessed 26April2023).

APPENDIX

All results in this article were collected with a Dell Technologies PowerEdge XE9680 server. The server was configured with 8 x 80 GB H100 NVIDIA® HGX GPU cards with NVLink™ technology, 2 TB of CPU RAM, and 2x Intel® Xeon® processors. The server was configured bare metal with the Ubuntu v22.04 Linux operating system, Anaconda v23.1.0, CUDA v12.1, cuDNN v8.8.1, and the same python dependencies as required by the HuggingFace Diffusers package.[12] The python dependencies included diffusers v0.14.0 and torch v2.0.0+cu118.

The image generation code was executed in the Jupyter Notebook (IPYNB) environment. The benchmarks were conducted in triplicate, but there was less than 1% deviation, thus error bars are not shown in the plots because they were too small to be seen. Benchmarks were conducted on a single GPU as multi-GPU image generation was not available at the time this article was written with the HuggingFace Diffusers library, float16 precision, and either the batch size of the model (bs = {1, 2, 4, 8, 16, 32, 64, 128, 256, 512}) or the n x n image resolution (n = {256, 512, 1,024, 2,048}) were altered. Only one variable was altered at a time. The model latency values, as runtime in seconds, were reported as outputs.

As an example, the following python3 IPYNB commands were used for benchmarking the HuggingFace Diffusers text-to-image single image generation models with 8 GPUs, batch size=1, float16 precision, number of iterations = 25, and 512 x 512 image resolution:

```
from diffusers import StableDiffusionPipeline, DPMSolverMultistepScheduler

model_id = “stabilityai/stable-diffusion-2-1”

# Using the DPMSolverMultistepScheduler (DPM-Solver++) scheduler
pipe = StableDiffusionPipeline.from_pretrained(model_id,
torch_dtype=torch.float16)
pipe.scheduler = DPMSolverMultistepScheduler.from_config(pipe.scheduler.config)
pipe = pipe.to(“cuda”)

prompt = “ Portrait of happy dog, close up”

image = pipe(prompt,
        height = 512,
        width = 512,
        num_inference_steps=25,
        guidance_scale=7.5, #default
        num_images_per_prompt=1,
        ).images[0]
```

The comparison of the Lambda image generation benchmarks and the performance of the Dell PowerEdge XE9680 server were conducted using the code provided by Lambda, as well as the dependencies described in their benchmark study.[14] The comparison to their benchmarks was performed with a single NVIDIA® H100 GPU HGX with 80 GB GPU RAM in the Dell PowerEdge XE9680 server.