FAST TRACK GENERATIVE AI WITH DELL™ POWEREDGE™ XE9680

MADE POSSIBLE WITH NVIDIA® H100 TENSOR CORE GPUS & BROADCOM 100 GIG-ETHERNET

SCALERS AI

Introduction

Unlock the potential of AI and transform your business with Dell™ PowerEdge™ XE9680 server - A machine that will elevate your AI performance to new heights.

The field of generative AI is experiencing an explosive growth, with cutting-edge developments in image, video, and audio media creation revolutionizing the creative industry. This remarkable technology is driving innovation in diverse sectors and opening up new frontiers for creative expression. Moreover, the immense promise of fine-tuned enterprise LLMs is bringing unparalleled insights to businesses, enabling them to protect their proprietary information, comply with data sovereignty issues, and improve their internal and external effectiveness. With the potential to process vast amounts of data in real-time, these finely-tuned models offer a decisive advantage in the fast-paced world of modern business. So, if you want to unlock the full potential of your data assets and stay ahead of the competition, enterprise LLMs are the way to go!

PERFORMANCE

Dell™ PowerEdge™ XE9680 Server (H100 vs A100)

1.8x

While Training GPT-2

3.15x

with the NVIDIA® Transformer engine enabled (Float8)

Dell™ PowerEdge™ XE9680 server with 8 NVIDIA® H100 GPUs and Broadcom BCM57508 NetXtreme-E 100G Ethernet offers unmatched performance for high-performance AI training and inference. Broadcom BCM957508-P2100G is a dual-port 100 Gb/s PCI Express 4.0 x16 Network Interface Card that supports QSFP56/QSFP28 optical modules and copper direct-attach cables which makes the card a perfect choice for network-intensive AI applications. Our whitepaper evaluates its effectiveness on common AI workloads like language and image recognition. Optimize your AI workloads and stay ahead of the competition with Dell™ PowerEdge™ XE9680.

Dell™ PowerEdge™ XE9680 is purpose-built for generative AI workloads, including language and generative AI models. We found it to demonstrate compelling performance in both training and fine-tuning of LLMs.

- Chetan Gadgil, CTO at Scalers AI™

The Results | Performance Summary (Training of GPT-2)

Dell™ PowerEdge™ XE9680 server with 8 NVIDIA **A100** GPUs and Broadcom BCM57508

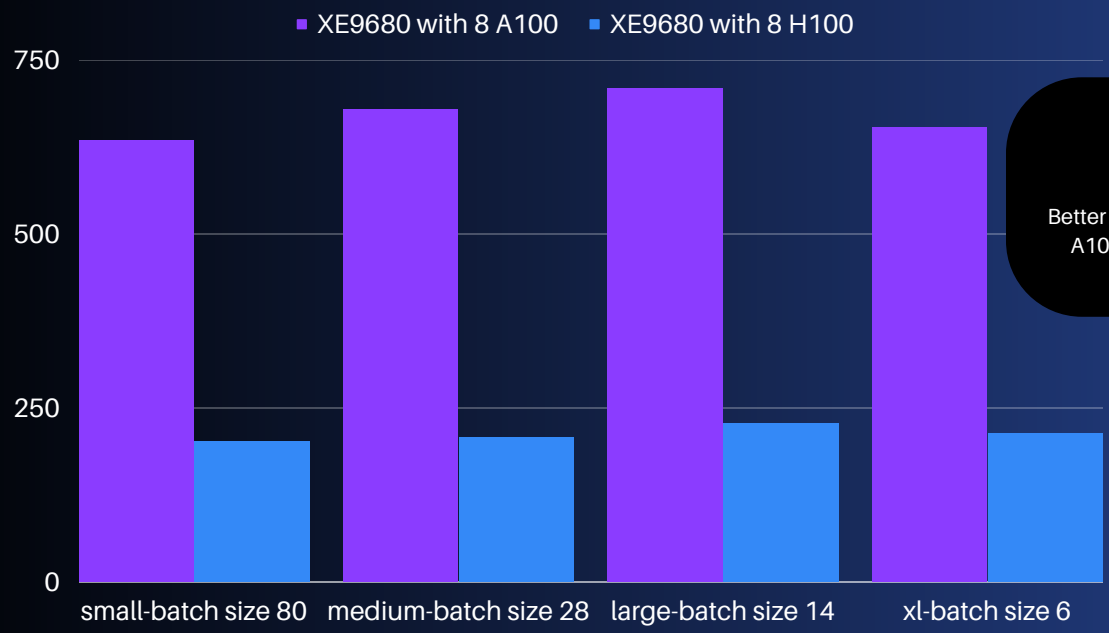
Dell™ PowerEdge™ XE9680 server with 8 NVIDIA **H100** GPUs and Broadcom BCM57508

1.79x

Better than A100 on H100 with gpt2-medium float32 precision

GPT-2 float32 - Training Step Time (ms)

Lower is Better



3.15x

Better than float32 precision on A100 on H100⁎ with float8 precision

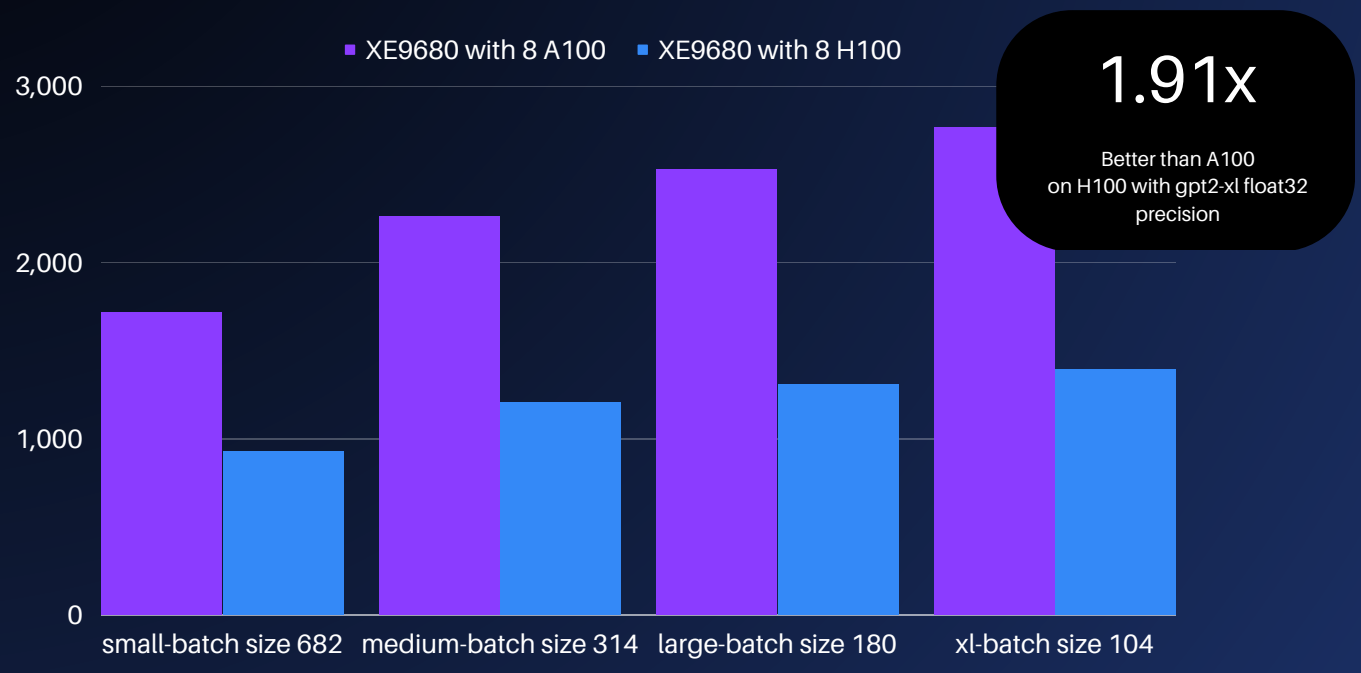
GPT-2 float8 - Training Step Time (ms)

With FP8 on Dell™ PowerEdge™ XE9680 server with 8 NVIDIA® H100 GPUs and Broadcom BCM57508

⁎Accuracy of FP8 models was not tracked. FP8 using the NVIDIA® Transformer Engine is available only on H100

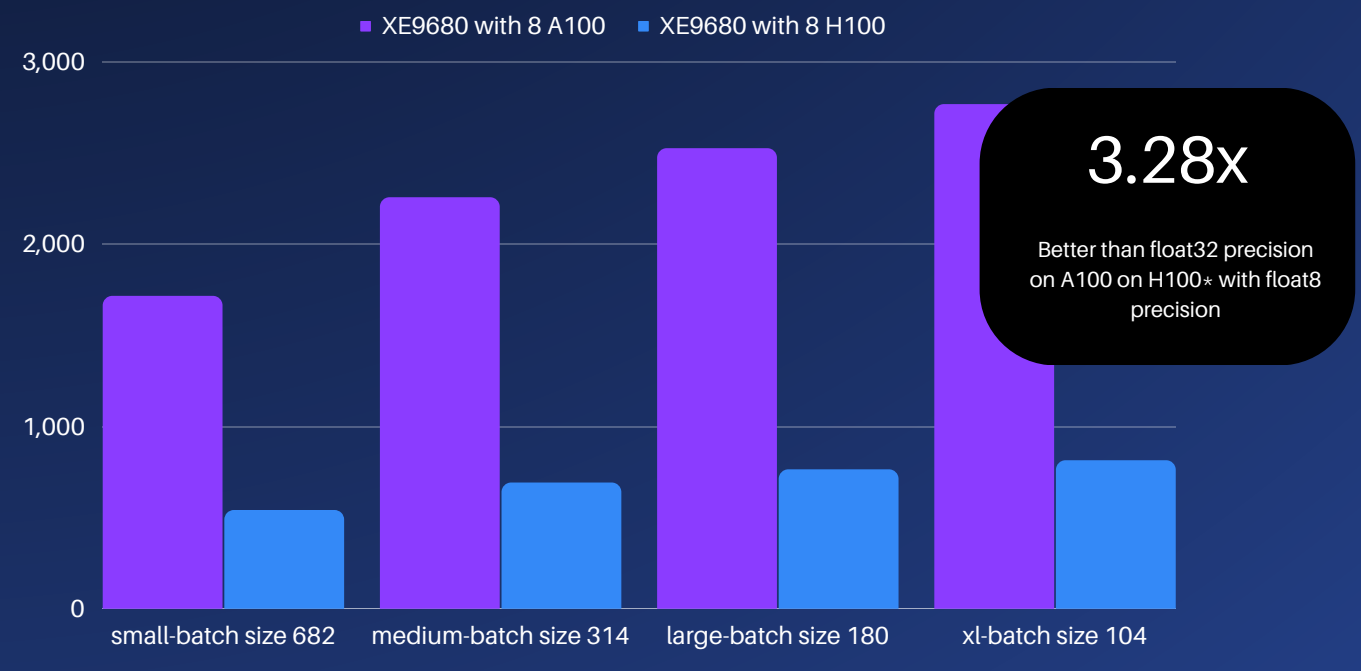
The Results | Performance Summary (Inference of GPT-2)

Dell™ PowerEdge™ XE9680 server with 8 NVIDIA® H100 GPUs and Broadcom BCM57508



GPT-2 float32 - Inference Step Time (ms) *Lower is Better*

With FP8 on Dell™ PowerEdge™ XE9680 server with 8 NVIDIA® H100 GPUs and Broadcom BCM57508



GPT-2 float8 - Inference Step Time (ms) *Lower is Better*

⁎Accuracy of FP8 models was not tracked. FP8 is available only on H100

⁎Disclaimer - Performance varies by use case, model, application, hardware & software configurations, the quality of the resolution of the input data, and other factors. This performance testing is intended for informational purposes and not intended to be a guarantee of actual performance of an AI application.

Example Real World Applications

Custom AI Chatbot

The Dell™ PowerEdge™ XE9680 offering for Generative AI workloads, specifically Large Language Models are very well suited. On the Dell™ PowerEdge™ XE9680 server, we successfully built, deployed, and operated a chatbot designed to answer queries about Dell Technology World 2023. Our approach began by generating embeddings using a pre-trained Large Language Model (LLM) that processed text from Dell Technology World 2023 website. These embeddings were then loaded into a vector database, forming a solid foundation for the chatbot's knowledge.

Next, we developed a micro-service specifically designed to handle question-and-answer interactions. By integrating this service with the vector database, we ensured the chatbot could efficiently retrieve relevant responses to user queries. To facilitate seamless user interaction, we created a user-friendly interface for the chatbot, enabling users to ask questions and receive answers with ease.

Finally, we deployed and executed the chatbot on Dell™ PowerEdge™ XE9680 H100 server, taking full advantage of its exceptional performance and capabilities. As a result, users could effectively engage with the chatbot to gather valuable information about Dell Technology World 2023, showcasing the server's versatility and power in real-world applications.

Stable Diffusion v2

Harnessing the impressive capabilities of Dell™ PowerEdge™ XE9680 H100 server, we successfully deployed and executed a pre-trained Stable Diffusion v2 model to generate intriguing images based on text prompts. The H100 server's remarkable performance and advanced features facilitated the efficient handling of the computationally demanding tasks inherent in the Stable Diffusion v2 model.

We deployed the pre-trained model onto Dell™ PowerEdge™ XE9680 H100 server, ensuring that it could fully utilize the server's processing power for real-time image generation. With the model in place, users could provide text prompts, prompting the Stable Diffusion v2 model, running on the H100 server, to generate visually appealing and contextually relevant images in response.

This process effectively demonstrated the server's ability to manage complex tasks and large-scale computations, highlighting its value for both creative and technical applications in the field of AI-driven image generation.

⁎Disclaimer - Performance varies by use case, model, application, hardware & software configurations, the quality of the resolution of the input data, and other factors. This performance testing is intended for informational purposes and not intended to be a guarantee of actual performance of an AI application.

EXAMPLE OUTPUT

PROMPT

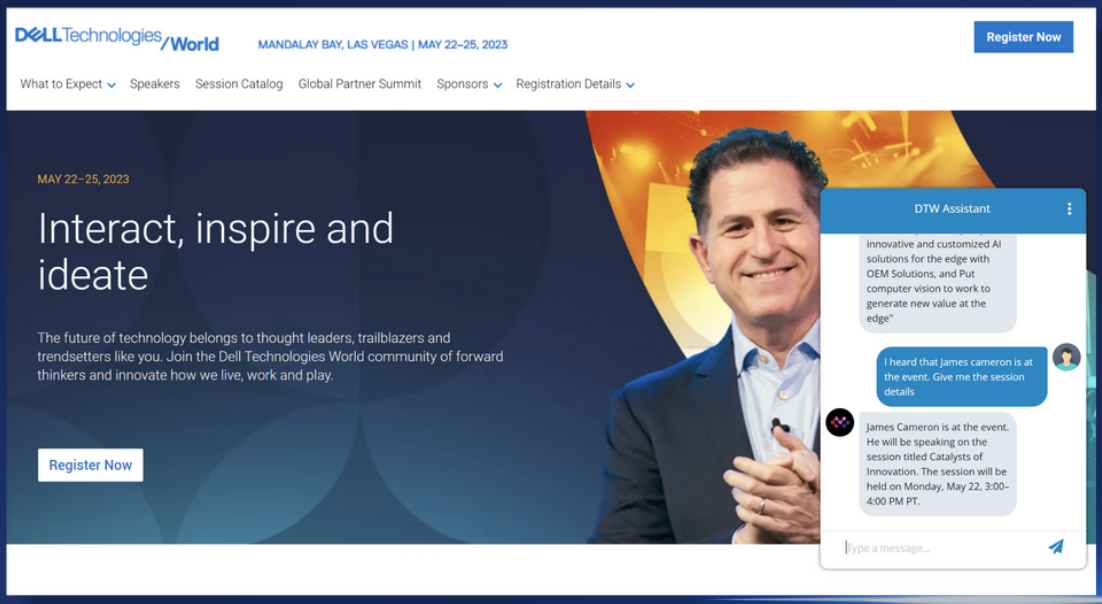
"a photo of an astronaut with red shirt riding a horse on moon"

OUTPUT

(Stable Diffusion 2)

PROMPT

"future city with artificial general intelligence"



Custom LLM chatbot running on Dell™ PowerEdge™ XE 9680
(powered by NVIDIA® H100)

Use Cases and Benefits

Dell™ PowerEdge™ XE9680 server has several potential use cases in various industries, including healthcare, finance, and retail. The improved performance and energy efficiency of the server can enable businesses to train AI models more quickly and accurately, leading to better predictions and insights. Dell™ PowerEdge™ XE9680 can also help businesses reduce their operational costs by enabling them to use less energy to perform AI workloads.

Enterprise Knowledge base

Fined tuned Language Models (LLMs) are an incredibly potent tool for enterprises to safeguard their proprietary information, adhere to data sovereignty issues, and enhance their internal and external (customer) effectiveness. The targeted fine-tuning with internal data empowers organizations to unlock the full potential of their data assets, gaining unprecedented insights into their business operations. With the ability to rapidly process vast amounts of data, fined tuned LLMs offer a decisive competitive edge in the dynamic business environment of today. So, if you are looking to remain ahead of the curve and leverage the full potential of your data assets, fined tuned LLMs are the way to go.

Dell™ PowerEdge™ XE9680 is a great platform to fine tune LLMs with your private and proprietary enterprise information rather than rely on bespoke, generic public APIs.

Generative AI

Generative AI is a fascinating technology that has revolutionized the field of artificial intelligence. Through its remarkable ability to generate realistic images and coherent text, it can create entirely new worlds that seem to come alive before your very eyes. This cutting-edge technology has the power to stimulate your imagination and spark your creativity, transporting you to places you never thought possible. From generating lifelike portraits to producing complex narratives, generative AI has ushered in a new era of innovation and discovery. Its ability to create novel and unique content has captivated the attention of researchers, scientists, and artists alike, and its influence on the future of AI is truly unparalleled.

Conclusion

The Dell™ PowerEdge™ XE9680 is well suited for Generative AI workloads, specifically LLM training, inference, and fine tuning. Whether it's the 8 A100 or 8 H100 config enterprises can dramatically reduce their time to industry transformation with this offering.

The H100 systems presents fantastic improvement with the Dell™ PowerEdge™ XE9680 server with 8 NVIDIA® H100 GPUs and Broadcom BCM57508 NetXtreme-E 100G Ethernet on an average being ~2x faster than the Dell™ PowerEdge™ XE9680 server with 8 NVIDIA® A100 GPUs and Broadcom BCM57508 NetXtreme-E 100G Ethernet system for training and inference large language models. Further, with the NVidia Transformer engine enabled oin Float8 (only supported in H100) we say an even higher performance increase to 3x on the latest

The PEX89000 family (ATLAS 2) of PCIe Gen 5.0 (32 GT/s) switches allows customers to build systems from simple PCIe connectivity inside the box to high-performance, low-latency, scalable, cost-effective PCIe fabrics for composable hyper-scale compute systems supporting ML/AI and Server/Storage applications.

About Scalers AI™

Scalers AI™ specializes in creating end-to-end artificial intelligence (AI) solutions for a wide range of industries, including retail, smart cities, manufacturing, and healthcare. The company is dedicated to helping organizations leverage the power of AI for their digital transformation. Scalers AI™ has a team of experienced AI developers and data scientists who are skilled in creating custom AI solutions for a variety of use cases, including predictive analytics, chatbots, image and speech recognition, and natural language processing. As a full stack AI solutions company with solutions ranging from the cloud to the edge, our customers often need versatile common off the shelf (COTS) hardware that works well across a range of workloads. Additionally, we also need advanced visualization libraries including the ability to render video in modern web application architectures.

Fast track development with access to the solution code

Save hundreds of hours of development with the solution code.
As part of this effort Scalers AI™ is making the solution code available.

Reach out to your Dell™ representative or contact Scalers AI™ at contact@scalers.ai for access.