DRAFT - Performance Insights

DELL™ POWEREDGE™ XE9680 SERVER WITH NVIDIA® H100 & A100 TENSOR CORE GPUS & BROADCOM 100 GIG-ETHERNET

SCALERS AI

Dell™ PowerEdge™ XE9680 with NVIDIA® H100 delivers >1.8x training & inference performance across GPT workloads and >3x with FP8 transformer optimizations enabled relative to NVIDIA® A100 system configurations.

Dell™ PowerEdge™ XE9680 Server with NVIDIA® H100 Tensor Core GPUs performance insights results in about 2x training & inference performance relative to Dell™ PowerEdge™ XE9680 Server with NVIDIA® A100 Tensor Core GPUs for Generative AI workloads using the GPT-2 Transformer Language (LLM) Model. By enabling FP8 transformer optimized on NVIDIA® H100 systems they delivered another boost to more than 3x NVIDIA® A100 systems.

Scalers AI™ conducted several test scenarios on Dell™ XE9680 Systems for modern Generative AI using GPT-2 Language Model Variants.

| Background

Transformer neural networks stand out for their unique ability to process input data in parallel rather than sequentially, making them exceptionally efficient in handling long-range dependencies within sequences. Their innovative self-attention mechanism allows them to focus on specific parts of the input while also considering the relationships between different elements, resulting in improved context-awareness and enhanced language understanding capabilities. Transformer models deliver exceptional understanding of complex data. These sophisticated models captivate the world of artificial intelligence, demonstrating the harmonious synergy between technology and human ingenuity. As businesses harness the potential of transformer networks, they will unlock new opportunities, paving the way for innovative solutions and strategic decision-making.
The unparalleled capabilities of these networks continue to shape the landscape of AI, fostering growth and driving success across diverse industries.

For the purpose of this evaluation, we leveraged the following models.

| Transformer Language Model Variants Evaluated: GPT-2 Variants

- **GPT2-small**
  - 12-layer, 768-hidden, 12-heads, 117M parameters
- **GPT2-medium**
  - 24-layer, 1024-hidden, 16-heads, 345M parameters
- **GPT2-large**
  - 36-layer, 1280-hidden, 20-heads, 774M parameters
- **GPT2-xl**
  - 48-layer, 1600-hidden, 25-heads, 1558M parameters

TRAINING DATASET

"superbench" generated random dataset

| Importance of Modeling Real-World Scenarios

It is crucial to consider the average performance across various workloads to gain a comprehensive understanding of a system's real-world performance because AI training and inference tasks encompass a diverse range of models, optimizations, and use cases. By evaluating the system across different workloads, it is possible to account for the variability and complexities present in real-life scenarios. This approach offers a more accurate representation of the system's capabilities, enabling developers and businesses to make informed decisions regarding its deployment and optimization. Moreover, examining the average performance across workloads ensures that the system is not overly optimized for a specific task at the expense of others, resulting in a more balanced and efficient system for diverse applications. Furthermore, the NVIDIA® Transformer Engine v0.7.0, which was utilized for training, was not accessible on the A100 system. Consequently, when evaluating "float8" scenarios on NVIDIA® H100 system, we resorted to using the optimal performance numbers from NVIDIA® A100 system as a basis for comparison with the float8 performance on NVIDIA® H100 system. This approach ensures a fair and comprehensive assessment of real-world performance across diverse workloads, offering valuable insights into the capabilities of both systems.

| NVIDIA® Transformer Engine Optimizations

We also evaluated the performance of NVIDIA® H100 system using the NVIDIA® transformer engine. The NVIDIA® Transformer Engine is a library for accelerating Transformer models on NVIDIA® GPUs, including using 8-bit floating point (FP8) precision on NVIDIA® H100 GPUs, to provide better performance with lower memory utilization in both training and inference1. By reducing the math to just eight bits, Transformer Engine makes it possible to train larger networks faster

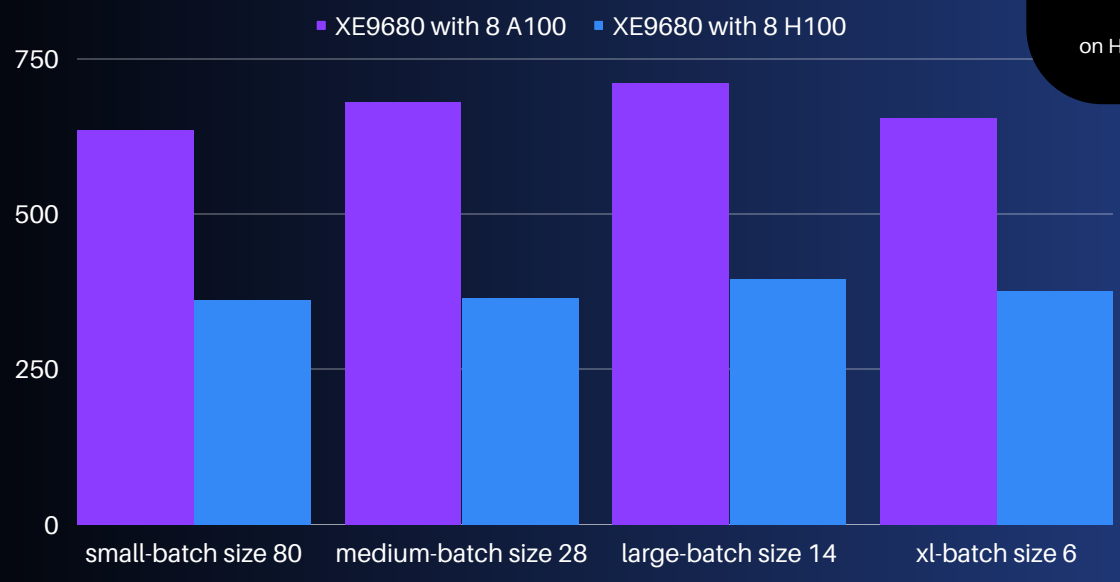
The Results | Performance Summary (Training of GPT-2)

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

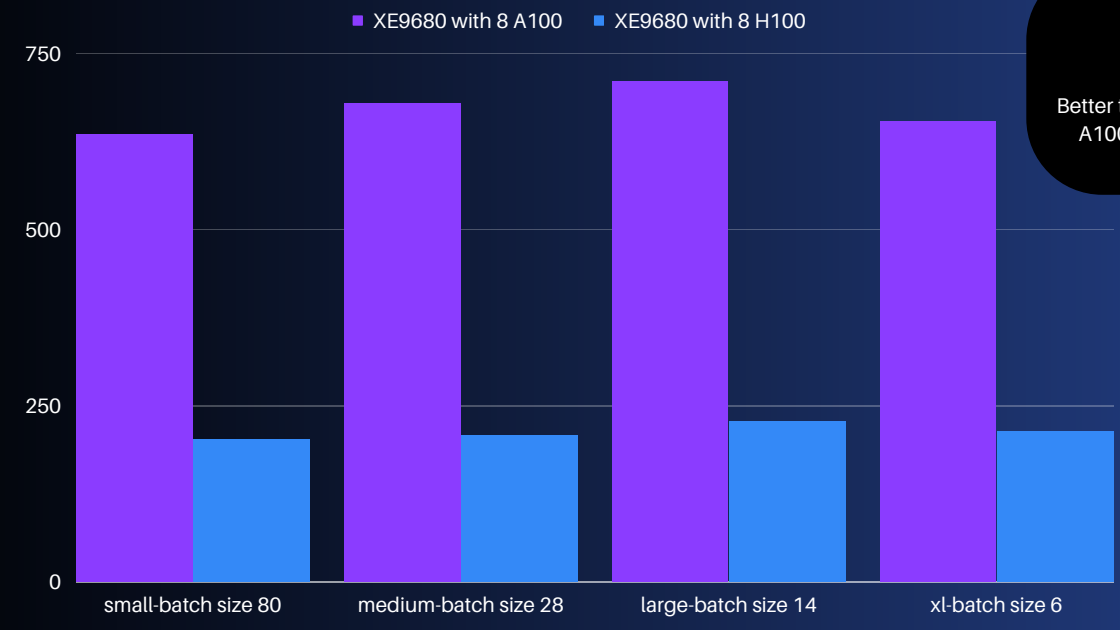
Better than A100 on H100 with gpt2-medium float32 precision



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

3.15x

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



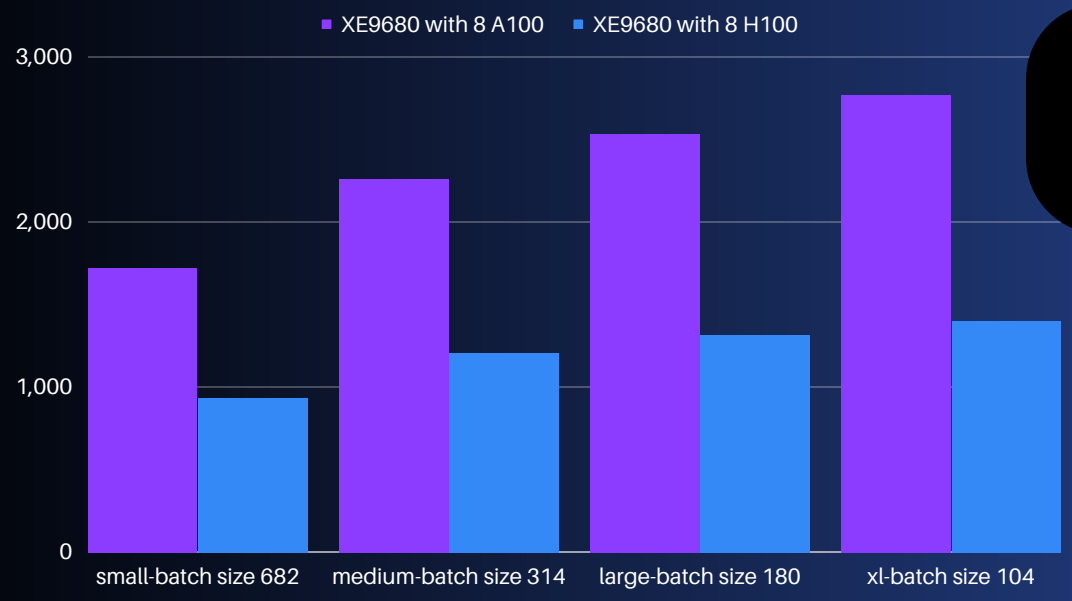
GPT-2 float8 for H100 & FP32 for A100- Training Step Time (ms)

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

⁎Accuracy of FP8 models was not tracked. FP8 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

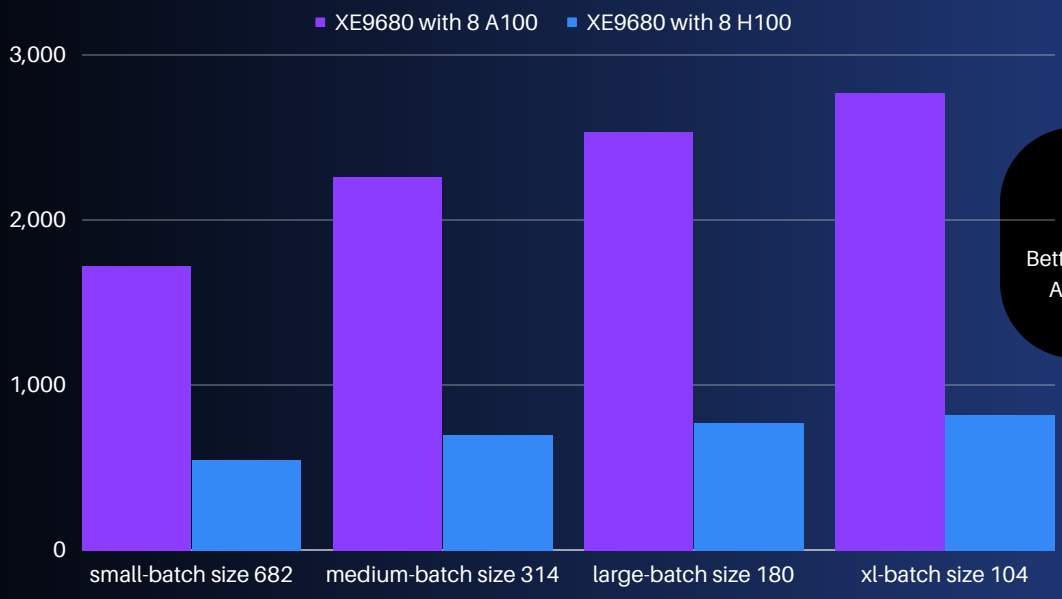


1.91x

Better than A100 on H100 with gpt2-xl float32 precision

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



3.28x

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

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

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

The following are our **Training** Results comparing the performance on systems with NVIDIA® A100 GPUs vs NVIDIA® H100 GPUs. Step times are in milliseconds (lower is better). The average improvement of the inference performance on the Dell™ PowerEdge™ XE9680 server with 8 NVIDIA® **H100** GPUs is **1.8x** compared to the Dell™ PowerEdge™ XE9680 server with 8 NVIDIA® **A100** GPUs.

- **Model: GPT2-large**
  - Precision: **Float32 ;** Batch size : 14
    - The A100 System: **709.60** ms step time
    - The H100 System: **393.60** ms step time
- **Model: GPT2-medium**
  - Precision: **Float32 ;** Batch size : **28**
    - The A100 System: **678.43** ms step time
    - The H100 System: **364.43** ms step time
- **Model: GPT2-small**
  - Precision: **Float32 ;** Batch size : **80**
    - The A100 System: **634.37** ms step time
    - The H100 System: **361.10** ms step time
- **Model: GPT2-xl**
  - Precision: **Float32 ;** Batch size : **6**
    - **The A100 System**: **652.79** ms step time
    - The H100 System: **374.51** ms step time

In addition, when switching to the **FP8** precision format (only available with NVIDIA® H100), we observed an average performance boost >3**x** compared to a similar workload on NVIDIA® A100.

Note: Model Accuracy was not compared between the A100 and the FP8 version on NVIDIA® H100.

The following are our **Inference** Results comparing the performance on systems with NVIDIA® H100 GPUs vs NVIDIA® A100 GPUs. Inference step time is in ms (lower is better) and Throughput is in inferences/sec (higher is better).
The average improvement of the inference performance on NVIDIA® H100 System is **1.9x** compared to NVIDIA® A100 System.

- **Model: GPT2-large**
  - Precision: **Float32 ,** Batch size : **180**
    - The A100 System: **2525.38** ms inference step time
    - The H100 System: **1308.34** ms inference step time
    - The A100 System: **19.74** inferences/sec throughput
    - The H100 System: **35.58** inferences/sec throughput
- **Model: GPT2-medium**
  - Precision: **Float32 ;** Batch size : **314**
    - The A100 System: **2256.21** ms inference step time
    - The H100 System: **1203.99** ms inference step time
    - The A100 System: **42.00** inferences/sec throughput
    - The H100 System: **76.95** inferences/sec throughput
- **Model: GPT2-small**
  - Precision: **Float32 ;** Batch size : **682**
    - The A100 System: **1715.02** ms inference step time
    - The H100 System: **927.04** ms inference step time
    - The A100 System: **127.47** inferences/sec throughput
    - The H100 System: **222.77** inferences/sec throughput
- **Model: GPT2-xl**
  - Precision: **Float32 ;** Batch size : **104**
    - The A100 System: **2767.30** ms inference step time
    - The H100 System: **1393.72** ms inference step time
    - The A100 System: **9.17** inferences/sec throughput
    - The H100 System: **16.01** inferences/sec throughput

⁎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.

The following are our **Training** Results highlighting the performance using the "**FP8_e4m3**" and "**FP8_hybrid**" precision on systems with NVIDIA® H100 GPUs vs NVIDIA® A100 GPUs which does not support the NVIDIA® Transformer Engine v0.7. Step times are in milliseconds (lower is better).

- **Model: GPT2-large**
  - Precision: FP8_e4m3 ; Batch size : 14
    - The H100 System: **227.03** ms step time
  - Precision: FP8_hybrid ; Batch size : 14
    - The H100 System: **227.34** ms step time
- **Model: GPT2-medium**
  - Precision: FP8_e4m3 ; Batch size : 28
    - The H100 System: **207.52** ms step time
  - Precision: FP8_hybrid ; Batch size : 28
    - The H100 System: **206.56** ms step time
- **Model: GPT2-small**
  - Precision: FP8_e4m3 ; Batch size : 80
    - The H100 System: **201.84** ms step time
  - Precision: FP8_hybrid ; Batch size : 80
    - The H100 System: **204.30** ms step time
- **Model: GPT2-xl**
  - Precision: FP8_e4m3 ; Batch size : 6
    - The H100 System: **213.11** ms step time
  - Precision: FP8_hybrid ; Batch size : 6
    - The H100 System: **213.04** ms step time

The following are our **Inference** Results highlighting the performance using the "FP8_e4m3" and "FP8_hybrid" precision on systems with NVIDIA® H100 GPUs vs NVIDIA® A100 GPUs which does not support the NVIDIA® Transformer Engine v0.7. Inference step times are in milliseconds (lower is better).

- **Model: GPT2-large**
  - Precision: FP8_e4m3 ; Batch size : **180**
    - The H100 System: **764.72** ms step time
  - Precision: FP8_hybrid ; Batch size : 14
    - The H100 System: **763.91** ms step time
- **Model: GPT2-medium**
  - Precision: FP8_e4m3 ; Batch size : **314**
    - The H100 System: **692.24** ms step time
  - Precision: FP8_hybrid ; Batch size : 80
  - The H100 System: **692.20** ms step time
- **Model: GPT2-small**
  - Precision: FP8_e4m3 ; Batch size : **682**
    - The H100 System: **541.45** ms step time
  - Precision: FP8_hybrid ; Batch size : 80
    - The H100 System: **541.06** ms step time
- **Model: GPT2-xl**
  - Precision: FP8_e4m3 ; Batch size : **104**
    - The H100 System: **814.59** ms step time
  - Precision: FP8_hybrid ; Batch size : 6
    - The H100 System: **815.25** ms step time

⁎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.

| Hardware & System Configurations

H100 SYSTEM

- System name: Dell PowerEdge XE9680 H100
- Motherboard: Dell Inc Model: 0F82N3 Version: X20
- CPU:
  - Model Name: Intel(R) Xeon(R) Platinum 8468
  - SpeedMax (MHz): 4000 MHz
  - Number Of Cores: 48
  - Number Of Sockets: 2
  - Thread(s) per core: 2
  - HTT: Yes
- GPU:
  - Model: NVIDIA Hopper (H100)
  - Count: 8
  - Memory: 80 GB
  - Max Power: 700 W
  - Type: SXM5
  - GPU Firmware: 530.30.02
- RAM:
  - Type: DDR5
  - Speed: 4800 MT/s
  - Size: 1 TB
- Storage:
  - Size: 1.5 TB
- NIC:
  - Product: BCM57508 NetXtreme-E 10Gb/25Gb/40Gb/50Gb/100Gb/200Gb Ethernet
  - Vendor: Broadcom Inc. and subsidiaries
  - Capacity: 25Gbit/s
- OS:
  - Name: Ubuntu 22.04.2 LTS
  - Kernel: Linux 5.15.0-69-generic
- Software:
  - CUDA Version: 12.1
  - Driver Version: 530.30.02
  - Pytorch: 2.0.0a0+1767026
  - Superbench: 0.8.0

A100 SYSTEM

- System name: Dell PowerEdge XE9680 A100
- Motherboard: Dell Inc Model: 0F82N3 Version: X20
- CPU:
  - Model Name: Intel(R) Xeon(R) Platinum 8480+
  - SpeedMax (MHz): 4000 MHz
  - Number Of Cores: 56
  - Number Of Sockets: 2
  - Thread(s) per core: 2
    - HTT: Yes
- GPU:
  - Model: NVIDIA Ampere (A100)
  - Count: 8
  - Memory: 80 GB
  - Max Power: 500 W
  - Type: SXM4
  - GPU Firmware: 530.30.02
- RAM:
  - Type: DDR5
  - Speed: 4800 MT/s
  - Size: 2 TB
- Storage:
  - Size: 2.9 TB
- NIC:
  - Product: Ethernet Controller E810-XXV for SFP
  - Vendor: Intel Corporation
  - Capacity: 25Gbit/s
- OS:
  - Name: Ubuntu 22.04.2 LTS
  - Kernel: Linux 5.15.0-70-generic
- Software:
  - CUDA Version: 12.1
  - Driver Version: 530.30.02
  - Pytorch: 2.0.0a0+1767026
  - Superbench: 0.8.0

*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.

Methodology

To ensure accurate and unbiased results, we configured the workloads to exercise a sustained system load of >90% on both NVIDIA A100 and H100-based systems. We continuously monitored the system load to maintain a consistent performance range. In order to gain a comprehensive understanding of expected system performance, we measured performance across a spectrum of commonly deployed applications on these systems. Workloads were selected based on leading enterprise AI application trends, including popular legacy and upcoming computer vision and language models. Additionally, we profiled both inference and training workloads, including training from scratch as well as fine-tuning pre-trained models such as transformers. We varied the precision while evaluating the training/inference workloads to understand the improvements in AI acceleration optimizations.

Building on this foundation, we tested several common, real-world scenarios on NVIDIA A100 and H100 based systems, respectively.

Scenarios

In the spirit of evaluating the performance of the Dell™ PowerEdge™ XE9680 in real-life enterprise scenarios, we conducted tests using various models, including GPT2-small, GPT2-medium, GPT2-large, and GPT2-xl. Each model was tested individually to establish performance baselines. Moreover, different precisions such as float32 and float16, as well as various batch sizes, were tested to determine maximum throughput.
The collected data was analyzed, revealing significant performance differences between the NVIDIA A100 and H100-based systems across different variables.

This dataset can aid in optimizing the Dell™ PowerEdge™ XE9680 for specific applications, enhancing AI workflows, and ensuring optimal performance for future use cases.

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.

Get Additional Performance Data

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