EINPRESSWIRE

PT study finds the Dell PowerEdge XE9680 servers with NVIDIA H100 GPUs supports many chatbot users

ROUND ROCK, TX, UNITED STATES, May 13, 2025 /EINPresswire.com/ -- Enterprises of all types and sizes are seeking to harness the power of GenAI to meet customer service needs via chatbots that can solve problems without human intervention. Finding a hardware solution that can support the right number of simultaneous chatbot users for a specific use case is crucial to success. Principled Technologies (PT) tested a Dell PowerEdge XE9680 servers with H100 GPUs to show the kind of chatbot performance an organization might expect with this solution.

A Principled Technologies report: Hands-on testing. Real-world results.

Running your in-house chatbot using Llama 3.1 405B LLMs on Dell PowerEdge XE9680 servers with NVIDIA H100 GPUs

Using your organization's own in-house data in conjunction with large language models (LLMs) can give your GenAI chatbots the specific information they need to improve questions and answers for users interacting with them. In fact, 38 percent of executives in a recent Gartner poll cited customer experience as the reason they adopted GenAI at all[1]—so the more seamless the experience, the more value it provides for companies and chatbot users alike. Chatbots require powerful GPUs to adequately support large numbers of simultaneous users asking questions and to provide these users with fast, accurate responses.

When it's time to select servers to power these AI projects, it can be difficult to find data to help form an accurate plan. We used the PTChatterly service to showcase the AI chatbot performance of a Dell™ PowerEdge™ XE9680 server powered by eight NVIDIA® H100 SXM™ Tensor Core GPUs, each with 80 GB of memory. For testing, we used a very large LLM (Llama 3.1 405B) and FP8 precision, which requires very fast accelerators and ample memory. With this LLM, which we augmented with in-house data, we quantified the number of supported users engaging in longer conversations to mirror what real-world users would experience. We used sample in-house data in conjunction with the LLM to address organizations that want the precision of a very large LLM using FP8 precision.

We found that a single Dell PowerEdge XE9680 server with eight NVIDIA GPUs (with 640 GB of total GPU memory) could support 68 simultaneous chatbot users engaging in lengthy questions and answers. For a 60kW rack of six PowerEdge XE9680 servers, we estimate that such a solution could support 408 simultaneous users at a cost of approximately $8.3M over the next five years.

If your organization needs the precision of a very large LLM using your own data, the Dell PowerEdge XE9680 server with NVIDIA H100 GPUs can provide ample resources to power your in-house chatbot.

Support up to 68 simultaneous chatbot users

Support 408 simultaneous users in a full rack for just $8.3M over 5 years

Running your in-house chatbot using very large LLMs on Dell PowerEdge XE9680 servers with NVIDIA H100 GPUs May 2025

Running your in-house chatbot using Llama 3.1 405B LLMs on Dell PowerEdge XE9680 servers with NVIDIA H100 GPUs

According to the report, Principled Technologies "used the PTChatterly service to showcase the AI chatbot performance of a Dell™ PowerEdge XE9680 server powered by eight NVIDIA H100 SXM TensorCore GPUs, each with 80 GB of memory. For testing, we used a very large LLM (Llama 3.1 405B) and FP8 precision, which requires very fast accelerators and ample memory. With this LLM, which we augmented with in-house data, we quantified the number of supported users engaging in longer conversations to mirror what real-world users would experience. We used sample in-house data in conjunction with the LLM to address organizations that want the precision of a very large LLM using FP8 precision.”

In this testing, PT found that a single Dell PowerEdge XE9680 server with eight NVIDIA GPUs could support 68 simultaneous chatbot users engaging in lengthy questions and answers. The report goes on to note that “for a 60kW rack of six PowerEdge XE9680 servers, we estimate that such a solution could support 408 simultaneous users at a cost of approximately $8.3M over the next five years. If your organization needs the precision of a very large LLM using your own data, the Dell PowerEdge XE9680 server with NVIDIA H100 GPUs can provide ample resources to power your in-house chatbot.”

To learn more about PT testing on the Dell PowerEdge XE9680 server with NVIDIA H100 GPUs, read the full report at https://facts.pt/mFNE3Nh.

About Principled Technologies, Inc.
Principled Technologies, Inc. is the leading provider of technology marketing and learning & development services.

Principled Technologies, Inc. is located in Durham, North Carolina, USA. For more information, please visit www.principledtechnologies.com.

Sharon Horton
Principled Technologies, Inc.
press@principledtechnologies.com
Visit us on social media:
LinkedIn
Facebook
YouTube
X