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

Introduction

The popularity of AI chatbots across the broad spectrum of global organizations continues to grow. In fact, Gartner experts predict that the growth of embedded conversational AI in enterprise apps jumped from 5 percent in 2020 to over 40 percent in 2024,[2] a testament to its swift and ongoing adoption. One popular use case for conversational AI is the AI chatbot. These chatbots answer customer service questions, help users navigate internal resources, and more. For some organizations, general-purpose AI models can serve their purposes well. For most, it's worth investing in in-house, domain-specific models with their own data to improve the outcomes for their own business use cases.

In addition, organizations have to choose the specific LLMs and LLM sizes that best meet their needs. When those needs require better contextual understanding and stronger reasoning capabilities, a very large LLM, such as the Llama 3.1 405B model we used in this study, may be the right solution.

Using LLMs in conjunction with private data to fuel successful in-house AI chatbots presents new challenges. Because GenAI is resource-intensive, selecting an appropriate hardware solution is key to success. First, you must be able to support the needs of your user base. Second, you have to manage the cost side of the cost-benefit equation, ideally choosing solutions with the right costs in both acquisition and operation.

About Dell PowerEdge XE9680 servers

The 6U Dell PowerEdge XE9680 is an eight-way GPU platform built to handle the most demanding workloads. According to Dell, this server is "designed and optimized for AI training & HPC use cases like LLM recommendation engines, molecular dynamics, and genome sequencing."[3]

With eight NVIDIA H100 GPUs, the PowerEdge XE9680 offers top-shelf memory capacity and bandwidth, and up to 1.5TB shared coherent GPU memory to improve generative AI training performance.[4] To learn more about how the Dell PowerEdge XE9680 can support your GenAI chatbots and other demanding projects, visit http://www.delltechnologies.com/asset/en-us/products/servers/technical-support/poweredge-xe9680-spec-sheet.pdf.

About NVIDIA H100 SXM Tensor Core GPUs

The NVIDIA H100 SXM GPUs we tested feature fourth-generation Tensor Cores and a Transformer Engine with FP8 precision. With 80GB memory and 3.35TB/s memory bandwidth, the H100 SXM GPU offers fourth-generation NVLink®, which supports 900 gigabytes per second (GB/s) of GPU-to-GPU interconnect to power workloads such as conversational AI.[5]

To learn more about NVIDIA H100 GPUs, visit https://www.nvidia.com/en-us/data-center/h100/.

To that end, we tested the GenAI chatbot performance using a very large LLM on the Dell PowerEdge XE9680 server with NVIDIA H100 GPUs. Our AI testing measured the full conversational capacity of the solutions with multi-question conversations involving a mix of question and response lengths. This approach let us see how many simultaneous users participating in such conversations this solution could support. By contrast, many approaches simply attempt to get maximum tokens-per-second throughput and then estimate the number of users by saying each user might have some fixed number of tokens per second. While presenting the tokens-per-second metric can provide insight into chatbot performance, you can use it only to compare the relative performance of different GPUs. By measuring full conversations as we did with our very large model, we can provide an accurate account of what users would experience using the chatbot in the real world.

We used an FP8 quantized model, instead of the original BF16 format used for training, to improve performance and to fit the entire model on a single server. Using FP8 offers these features with very little change in model accuracy.

PTChatterly is a testing service that can help organizations size and understand a solution's performance for an in-house chatbot that utilizes RAG with a popular LLM and a private database of business information. In testing, we selected a number of variables, including the LLM, the corpus of data, the response time threshold, and the response time threshold percentile. We define the response time threshold as the longest acceptable amount of time the chatbot takes to fully answer a question—typically 5 to 30 seconds—and the response time threshold percentile as the proportion of questions that must complete below that threshold.

For this test, we used the following variables:

- **LLM:** The Llama 3.1 405B model, with 405 billion parameters. The LLM framework or library we used to run the model was vLLM. Meta describes this model as "the first openly available model that rivals the top AI models when it comes to state-of-the-art capabilities in general knowledge, steerability, math, tool use, and multilingual translation."[6]
- **Corpus of data:** We used a text-only corpus of Airbnb rental data that includes details about home listings and reviews from customers, from which we scrubbed any obvious personal information (e.g., host names) before ingesting. This corpus is a good representation of retail-style data because it includes product descriptions, pricing, and other information to help customers make decisions.
- **Response time threshold** and **response time threshold percentile:** We chose a median response time threshold of 15 seconds, and 95th percentile threshold of 30 seconds. This means the solution answered the majority of questions in their entirety in less than 15 seconds, and 95 percent of them in less than 30 seconds. In all our tests, the 95th percentile threshold was never reached (no responses took more than 30 seconds), and so we focus primarily on the median threshold for reporting, as the benchmark would always stop at the median threshold first. Note that the response time refers to the number of seconds is how long the complete answer takes to appear; the first characters appear in under 2 seconds, and within the total of 15 seconds, users would see a scrolling answer so they can read it as it completes.

Then, we calculated the total cost of ownership (TCO) of a rack of six servers over the next five years. To learn more about how we tested, read the science behind the report.

PTChatterly

PTChatterly helps you size and understand a solution's performance for an in-house chatbot that supplements a popular LLM with a private dataset. It couples a full-stack AI implementation of an LLM, augmented with in-house data, with a testing harness that lets you determine how many people the chatbot can support. Rather than reporting results in technical measurements that few users would understand, it provides a metric that is meaningful and simple to grasp: For example, it might say that the server under test supports 32 people having simultaneous conversations with a response time of 10 seconds or less. Learn more at PTChatterly.ai.

Support up to 68 simultaneous chatbot users on a Dell PowerEdge XE9680 server powered by NVIDIA H100 GPUs

Think about your ideal in-house chatbot and its planned use case. How many users do you expect to converse with your chatbot at a time? To help organizations plan for the number of servers they might need, we ran PTChatterly on the Dell PowerEdge XE9680 server with NVIDIA H100 GPUs and determined the expected user count it could support. As Figure 1 shows, this configuration can support up to 68 simultaneous chatbot users with response times below 15 seconds.

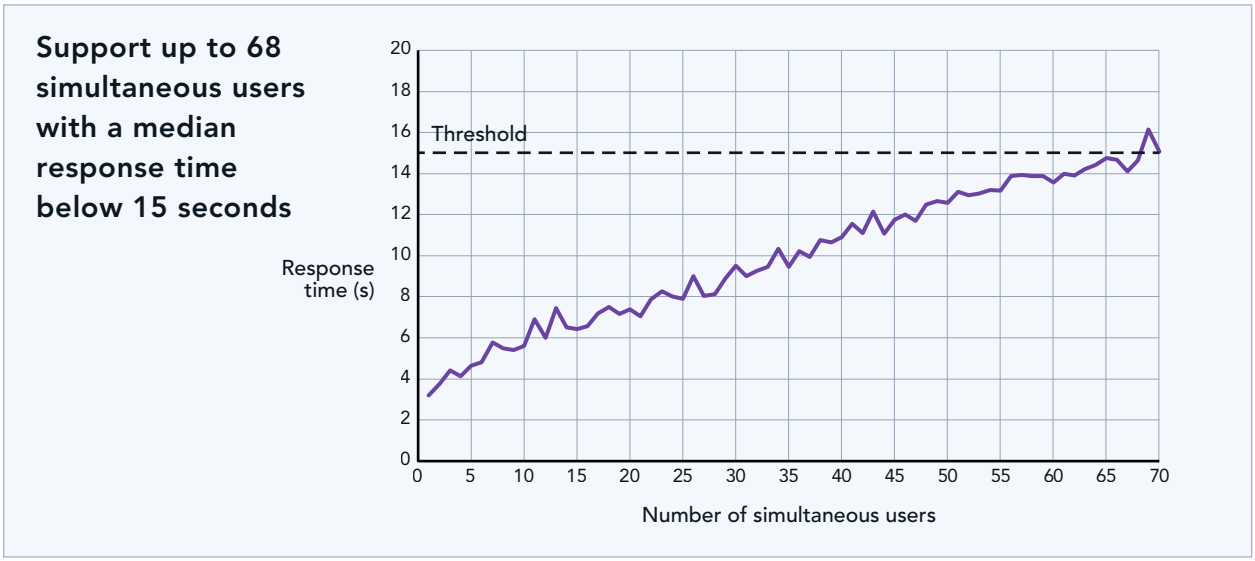

Figure 1: Number of simultaneous chatbot users that the Dell PowerEdge XE9680 supported with NVIDIA H100 GPUs. Higher is better. Source: Principled Technologies.

Need to support more chatbot users? Here's what performance for a 60kW rack of these servers could look like

Based on our testing with the entire server dedicated to GenAI, Figure 2 shows the extrapolation of simultaneous chatbot users that we expect Dell PowerEdge XE9680 servers could support in a 60kW rack configuration.

In our tests, the power utilization of the server peaked at 9kW. Consequently, we estimate that a 60kW rack can support up to six servers (six servers using 9 kW each yields 54 kW of usage). Each server can run one instance of the LLM, and all the servers can work in parallel, so we can extrapolate linear scaling of 408 simultaneous users in a rack (six servers supporting 68 simultaneous users each).

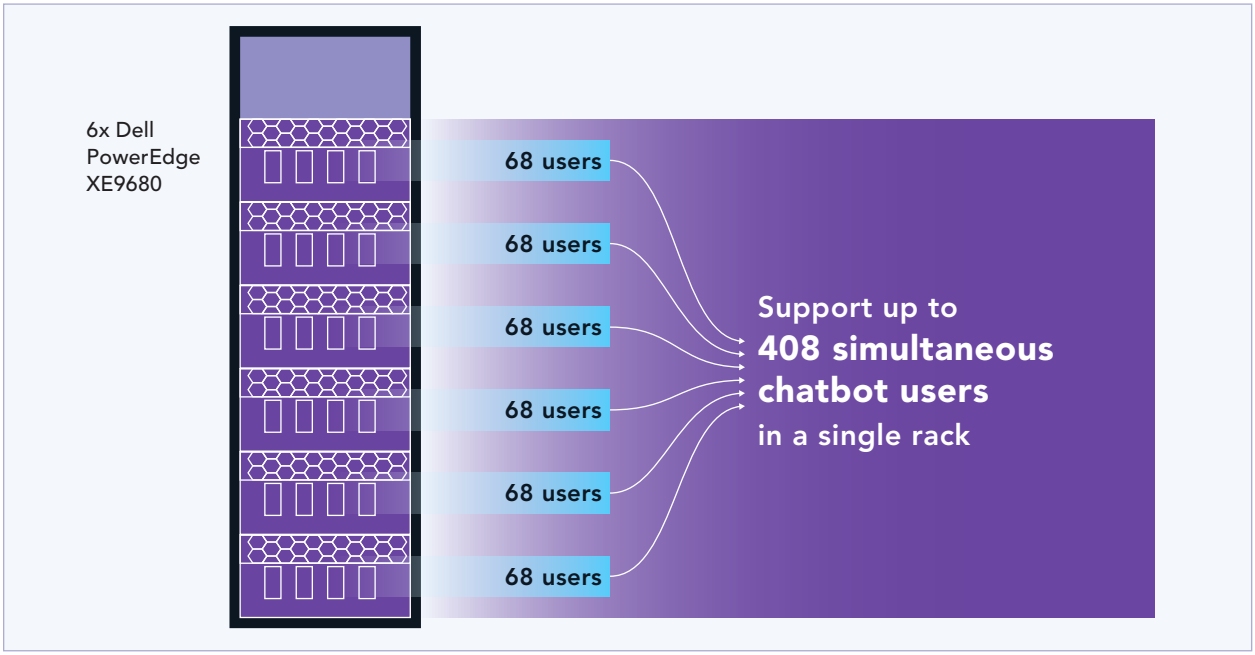

Figure 2: Extrapolation of the number of simultaneous chatbot users that you could expect for a 60kW rack of Dell PowerEdge XE9680 servers with NVIDIA H100 GPUs. Higher is better. Source: Principled Technologies.

Calculating the 5-year TCO for six Dell PowerEdge XE9680 servers, each with eight NVIDIA H100 GPUs

Organizations that plan to scale GenAI chatbots across multiple servers may have questions about what the total acquisition and operating costs could look like over the next 5 years. We ran the numbers to show the approximate 5-year costs you might expect for supporting an AI infrastructure comprising six of these servers.

Based on our calculations, we estimate a rack of six Dell PowerEdge XE9680 servers, each with eight NVIDIA H100 GPUs, would cost approximately $8.3 million over five years. As our PTChatterly testing showed, we assume that this rack would be able to support 408 simultaneous chatbot users having real-world conversations.

Table 1 breaks down our 5-year cost calculations for six Dell PowerEdge XE9680 servers supporting 408 users. For further details, including the specific assumptions we made to arrive at these calculations, see the science behind the report. Note: As with any TCO estimate, your cost savings will vary depending on several factors.

Table 1: TCO summary. Lower costs are better. Source: Principled Technologies.

5-year TCO summary for 6x Dell PowerEdge XE9680 servers, each with 8x NVIDIA H100 GPUs	
Number of systems required (per rack)	6
Total hardware cost (1 rack)	$8,062,352.94
Total 5y power cost	$253,372.52
Total 5y data center space cost	$10,000.00
Total 5y maintenance cost	$39,110.67
Total 5-year TCO	**$8,364,836.13**

- The acquisition costs we use in our comparison reflect pricing from Dell we received in December 2024. Discounts, taxes, and other fees will vary.
- To calculate the costs of power consumption, we used the active and idle wattage we measured in our testing, and calculated 5 years of power using the average energy cost in the US as of November 2024: $0.1701 per kWh.[7]
- Maintenance (administration) costs we use in our calculation include the assumption that a single IT admin is responsible for 100 servers, and divide the number of servers. We then multiplied this number by an average salary of $100,580 for a network and computer systems administrator to determine labor costs/admin burden.[8]

Conclusion

Supporting an in-house chatbot that uses your own corporate data with the precision of a very large LLM requires a powerful server-and-GPU solution. In our test case using the very large Llama 3.1 405B LLM, the Dell PowerEdge XE9680 server with eight NVIDIA H100 GPUs supported 68 simultaneous chatbot users. Based on our test results, we estimate that a rack of six Dell PowerEdge XE9680 servers, each with eight NVIDIA H100 SXM GPUs, would support 408 simultaneous users with a five-year TCO of approximately $8.3 million.

1. Gartner, "Gartner Experts Answer the Top Generative AI Questions for Your Enterprise," accessed February 3, 2025, https://www.gartner.com/en/topics/generative-ai.
2. Gartner, "Gartner Experts Answer the Top Generative AI Questions for Your Enterprise."
3. Dell, "PowerEdge XE9680 Specification Sheet," accessed April 11, 2025, http://www.delltechnologies.com/asset/en-us/products/servers/technical-support/poweredge-xe9680-spec-sheet.pdf.
4. Dell, "PowerEdge XE9680 Specification Sheet."
5. NVIDIA, "NVIDIA H100 Tensor Core GPU Datasheet," accessed April 11, 2025, https://resources.nvidia.com/en-us-tensor-core/nvidia-tensor-core-gpu-datasheet?ncid=no-ncid.
6. Meta, "Introducing Llama 3.1: Our most capable models to date," accessed April 11, 2025, https://ai.meta.com/blog/meta-llama-3-1/.
7. Average price of electricity, November 2024, https://www.eia.gov/electricity/monthly/epm_table_grapher.php?t=table_es1a.
8. Average for network and computer systems administrator, BLS May 2023, https://www.bls.gov/oes/current/oes_nat.htm.

Read the science behind this report at https://facts.pt/TGlBTr9 ▶

Principled Technologies®

Facts matter.®

This project was commissioned by Dell Technologies.