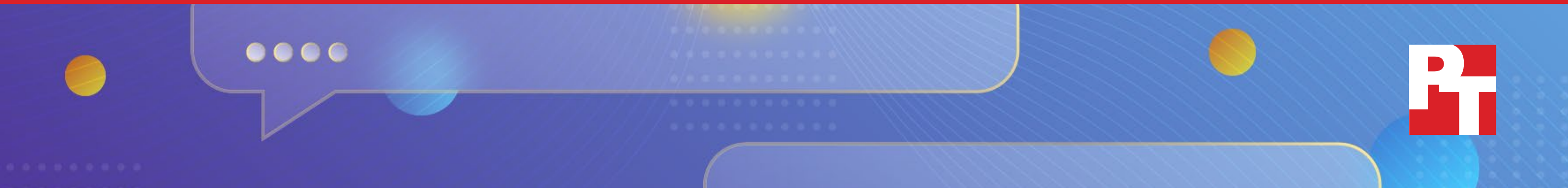
A Principled Technologies report: Hands-on testing. Real-world results.



Dell PowerEdge XE9680 servers with AMD Instinct MI300X Accelerators: the power to host GenAI with Llama 3.1 405B LLMs

As generative AI (GenAI) adoption continues, organizations are increasingly using their specific in-house data in conjunction with large language models (LLMs) to provide AI chatbots to meet customer needs. In a recent Gartner poll, 38% of executives cited customer experience as the primary purpose for using GenAI[1]—which means you need a hardware solution that's powerful enough to support the number of chatbot users you expect to serve and give them quick, accurate responses.

Organizations have many ways to configure servers to meet their AI needs, but specific data to help them plan may be scarce. To assist these businesses, we used the PTChatterly service to showcase the AI chatbot performance of a Dell™ PowerEdge™ XE9680 server powered by AMD Instinct™ MI300X Accelerators with an industry-leading 192 GB of high bandwidth memory (HBM3) in two different use cases: with eight accelerators and then using just four of the accelerators, for those who wish to use their accelerators for multiple workloads. In fact, the huge memory size of the AMD Instinct MI300X Accelerators is what makes it possible to run a very large LLM on only four accelerators; it has the most memory of any available GPU as of this writing. For testing, we used a very large Llama 3.1 405B LLM (which has 405 billion parameters) and FP8 precision, which requires very fast accelerators that have HBM. With this LLM, which we augmented with in-house data, we were able to quantify the number of supported users engaging in longer conversations to mirror what real-world users would experience.

Both Dell PowerEdge XE9680 server use cases supported substantial numbers of chatbot users: using four accelerators, the server supported 72 simultaneous users while leaving room for other workloads, while the standard eight accelerator-use case supported 136 simultaneous users. For organizations looking to support in-house chatbots using their own data and to use a high-precision very large LLM, these results show what's possible as you allocate resources for your new AI infrastructure. To help organizations understand how much a GenAI project might cost, we also calculated expected five-year TCO costs.

When your business requires the precision of a very large LLM, Dell PowerEdge XE9680 servers with AMD Instinct MI300X Accelerators offer the resources you need to make your in-house GenAI project a success.

Support up to 72 simultaneous chatbot users

using only 4 accelerators; the other 4 are free for other applications

Support up to 136 simultaneous chatbot users

with 8 accelerators

Support 816 simultaneous users in a full rack for just $7.6M

over 5 years

> *"By 2027, more than 50% of the GenAI models that enterprises use will be specific to either an industry or business function — up from approximately 1% in 2023."*
>
> *—Gartner*[2]

Introduction

An ever-increasing number of organizations are leveraging this tech to supercharge their business operations, using AI chatbots to answer customer service questions, help users navigate resources, and more. While general-purpose AI models may suffice for some purposes, they won't meet most business needs. Instead, organizations are working to create in-house, domain-specific models with their own data to improve the outcomes for their own business use cases. In fact, Gartner predicts that "By 2027, more than 50% of the GenAI models that enterprises use will be specific to either an industry or business function — up from approximately 1% in 2023."[3]

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, like 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.

To give organizations help in planning how to allocate resources for their specific GenAI project, we tested the performance of Dell PowerEdge XE9680 servers with AMD Instinct MI300X Accelerators two ways. First, we allocated just four accelerators for our GenAI workload, leaving the other four accelerators available to run other workloads, as some as some organizations might prefer. We next tested utilizing all eight accelerators for our GenAI workload.

About Dell PowerEdge XE9680 servers

The 6U Dell PowerEdge XE9680 is an 8-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."[4]

With 8 AMD Instinct MI300X 750W OAM accelerators fully interconnected with Infinity Fabric links, 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.[5] 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 AMD Instinct MI300X Accelerators

Based on 4th Gen Infinity Architecture, the AMD Instinct MI300X Accelerators are Open Accelerator Modules (OAM) with 8 Infinity Fabric™ links with up to 128 GB/s peak Infinity Fabric link bandwidth and are "designed to deliver leadership performance for Generative AI workloads and HPC applications."[6]

Per AMD's datasheet about the Instinct M1300X, "It is designed with 304 high-throughput compute units, AI-specific functions including new data-type support, photo and video decoding, plus an unprecedented 192 GB of HBM3 memory on a GPU accelerator."[7] To learn more about AMD Instinct M1300X Accelerators, visit https://www.amd.com/en/products/accelerators/instinct/mi300/mi300x.html.

To test the two use cases, we used the PTChatterly testing service to determine the number of simultaneous users of a chatbot augmented with in-house data that each could support. We used a very large LLM (Llama 3.1 405B with FP8 precision) to measure the performance an organization that wanted the precision of such a model could expect to experience. 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 each 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 accelerators. 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. AMD Instinct MI300X Accelerators have special support for FP8 acceleration, which allows them to achieve significantly faster performance when using this level of precision.

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 is called 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."[8]
- **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 this 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 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.

Choose the Dell PowerEdge XE9680 with AMD Instinct MI300X Accelerators to power your chatbots

We first ran PTChatterly on the Dell PowerEdge XE9680 server taxing only four of the AMD Instinct MI300X Accelerators to represent a use case where an organization wants to use four accelerators for GenAI and four for another application. Figure 1 shows that this use case could support up to 72 simultaneous chatbot users with response times below 15 seconds.

Support up to 72 simultaneous users with a median response time below 15 seconds

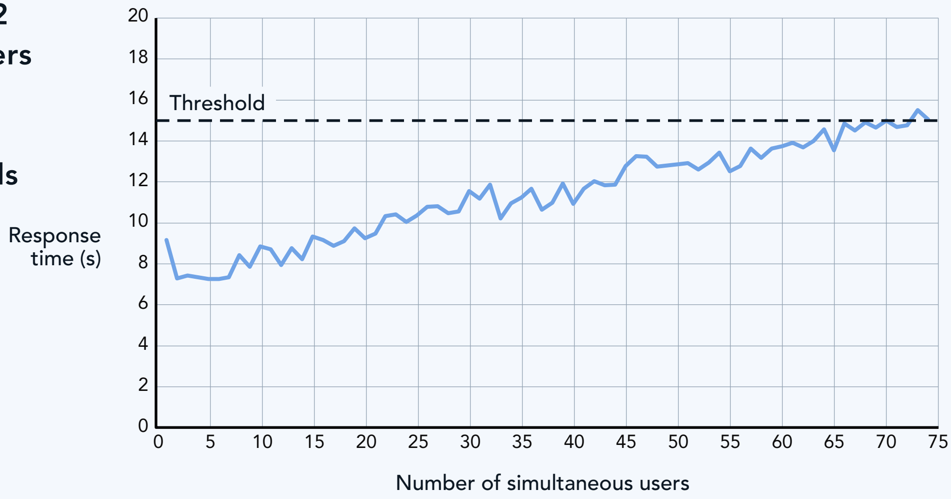


Figure 1: Number of simultaneous chatbot users that the Dell PowerEdge XE9680 supported using four accelerators. Higher is better. Source: Principled Technologies.

Then, we ran PTChatterly on the Dell PowerEdge XE9680 server using all eight AMD Instinct MI300X Accelerators to show performance for the server entirely dedicated to GenAI. (One instance of the LLM ran on each set of four accelerators.) Figure 2 shows that this use case scaled to support up to 136 simultaneous chatbot users with response times below 15 seconds.

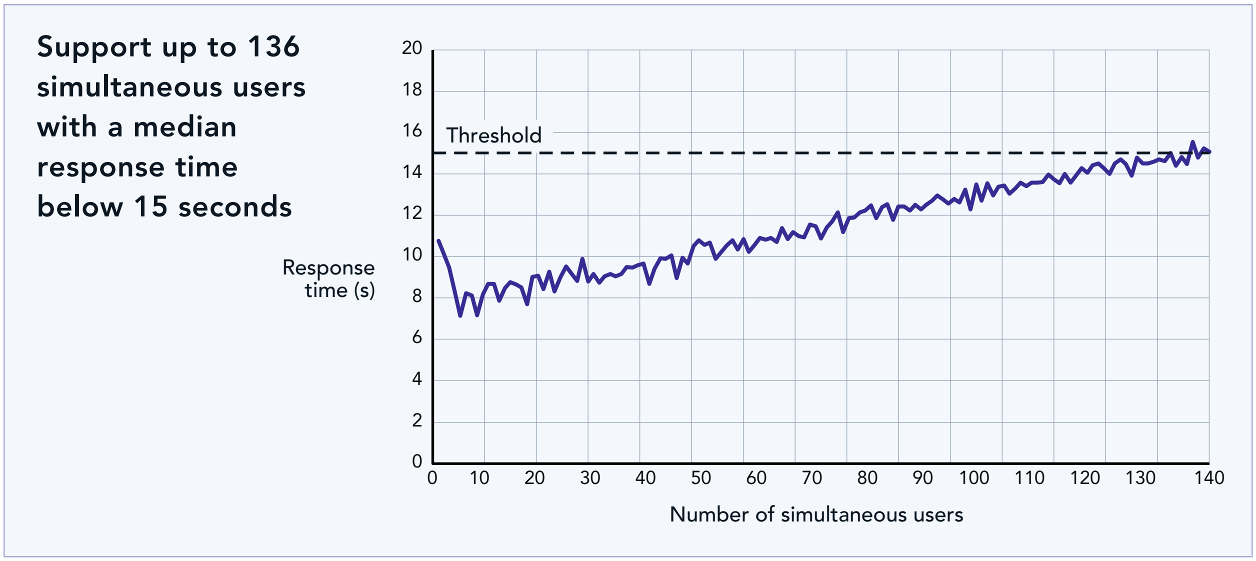


Figure 2: Number of simultaneous chatbot users that the Dell PowerEdge XE9680 supported using all eight accelerators. Higher is better. Source: Principled Technologies.

Support up to 816 simultaneous chatbot users in a single rack

Based on our testing with the entire server dedicated to GenAI, Figure 3 shows the extrapolation of simultaneous chatbot users that we expect the Dell PowerEdge XE9680 servers to 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 two instances of the LLM, and all the servers can work in parallel, so we can extrapolate linear scaling of 816 simultaneous users in a rack (six servers each delivering 136 simultaneous users).

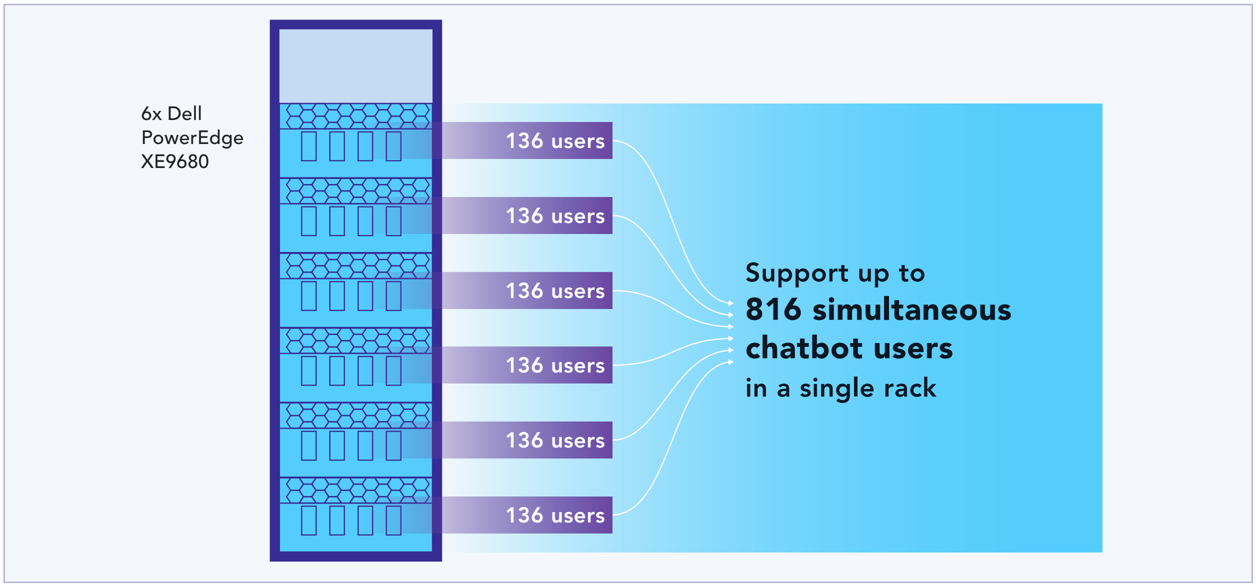


Figure 3: Extrapolation of the number of simultaneous chatbot users that you could expect for a 60kW rack of Dell PowerEdge XE9680 servers. Higher is better. Source: Principled Technologies.

Exploring the 5-year TCO for a rack of Dell PowerEdge XE9680 servers with eight AMD Instinct MI300X Accelerators

While maximum chatbot performance may be the goal for some organizations, others seek to balance strong performance while also leaving room to grow. If your organization plans to dedicate a full rack of servers to run your in-house chatbot, we ran the numbers to show the approximate five-year costs you might expect for supporting this robust AI infrastructure.

We estimate a full rack of Dell PowerEdge XE9680 servers, each with eight AMD Instinct MI300X Accelerators, would cost approximately $7.6 million over five years. As our PTChatterly testing showed, we assume that this full rack would be able to support 816 simultaneous chatbot users having real-world conversations.

Table 1 breaks down our 5-year cost calculations for six Dell PowerEdge XE9680 servers. 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 AMD Instinct MI300X Accelerators	
Number of systems required (per rack)	6
Total hardware cost (1 rack)	$7,280,228.94
Total 5y power cost	$291,168.30
Total 5y data center space cost	$10,000.00
Total 5y maintenance cost	$39,110.67
Total 5-year TCO	**$7,620,507.91**

- 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.[9]
- 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.[10]

Conclusion

When your goal is to have the precision of a very large LLM, as in our test case the Llama 3.1 405B LLM, supporting an in-house chatbot augmented with your corporate data, the Dell PowerEdge XE9680 server with AMD Instinct MI300X Accelerators is a strong choice. We found that using just half of the accelerators, the server supported 72 simultaneous users while leaving significant resources to run other workloads. Performance scaled well, with the full eight accelerators supporting up to 136 simultaneous users. Based on our test results, we estimate that supporting a rack of six Dell PowerEdge XE9680 servers each with eight AMD Instinct MI300X Accelerators would support 816 simultaneous users with a five-year TCO of approximately $7.6 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] Arun Chandrasekaran, "3 Bold and Actionable Predictions for the Future of GenAI," accessed January 31, 2025, https://www.gartner.com/en/articles/3-bold-and-actionable-predictions-for-the-future-of-genai.

[3] Arun Chandrasekaran, "3 Bold and Actionable Predictions for the Future of GenAI."

[4] Dell, "PowerEdge XE9680 Specification Sheet," accessed March 20, 2025, http://www.delltechnologies.com/asset/en-us/products/servers/technical-support/poweredge-xe9680-spec-sheet.pdf.

[5] Dell, "PowerEdge XE9680 Specification Sheet."

[6] AMD, "AMD Instinct MI300X Accelerators," accessed March 20, 2025, https://www.amd.com/en/products/accelerators/instinct/mi300/mi300x.html.

[7] AMD, "AMD Instinct MI300X Accelerator Datasheet," accessed March 20, 2025, https://www.amd.com/content/dam/amd/en/documents/instinct-tech-docs/data-sheets/amd-instinct-mi300x-data-sheet.pdf.

[8] Meta, "Introducing Llama 3.1: Our most capable models to date," accessed March 28, 2025, https://ai.meta.com/blog/meta-llama-3-1/.

[9] Average price of electricity, November 2024, https://www.eia.gov/electricity/monthly/epm_table_grapher.php?t=table_es1a.

[10] 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/0puJB4P ▶

Principled Technologies®

Facts matter.®

This project was commissioned by Dell Technologies.