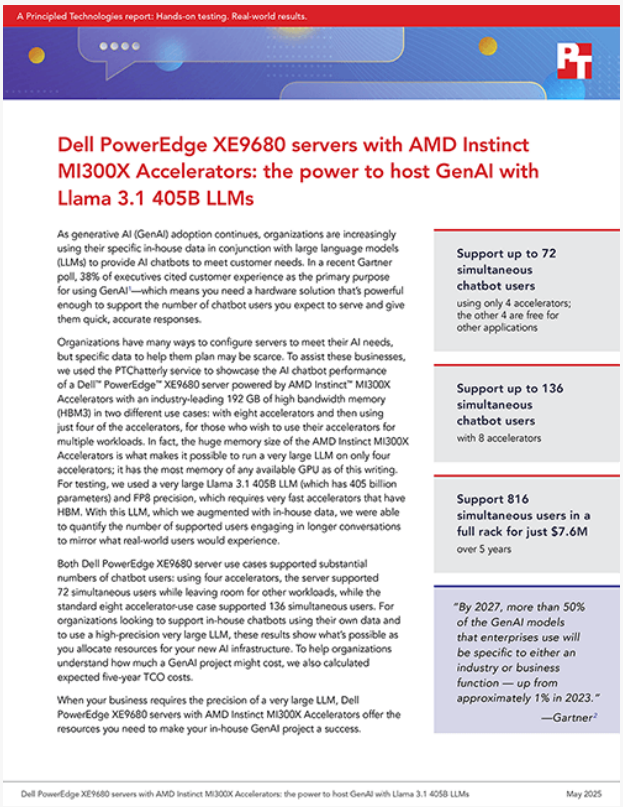
PT study shows that Dell PowerEdge XE9680 servers with AMD Instinct MI300X GPUs make a strong platform for GenAI

In hands-on testing, a Dell PowerEdge XE9680 server with AMD Instinct MI300X Accelerators supported up to 136 simultaneous chatbot users

ROUND ROCK, TX, UNITED STATES, May 13, 2025 /EINPresswire.com/ -- As more and more organizations are looking to use their own data to support GenAI chatbots to answer customer and employee questions, the need for real-world data about appropriate solutions increases. Principled Technologies (PT) tested a Dell PowerEdge XE9680 servers with AMD Instinct MI300X Accelerators to show the kind of chatbot performance an organization might expect.



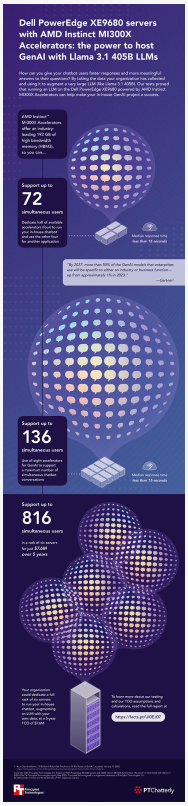
Dell PowerEdge XE9680 servers with AMD Instinct MI300X Accelerators: the power to host GenAI with Llama 3.1 405B LLMs

According to the report, "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."

PT used the very large Llama 3.1 405B LLM and FP8 precision, which requires very fast accelerators that have HBM. They augmented the LLM with in-house data and used the PTChatterly service to determine how many users could engage in long conversations. This approach mirrors what real-world LLM users would likely experience.

The report continues, "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.



Infographic

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."

To learn more, read the full report at https://facts.pt/ovSGH4f or see the infographic at https://facts.pt/6uHzsvR.

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