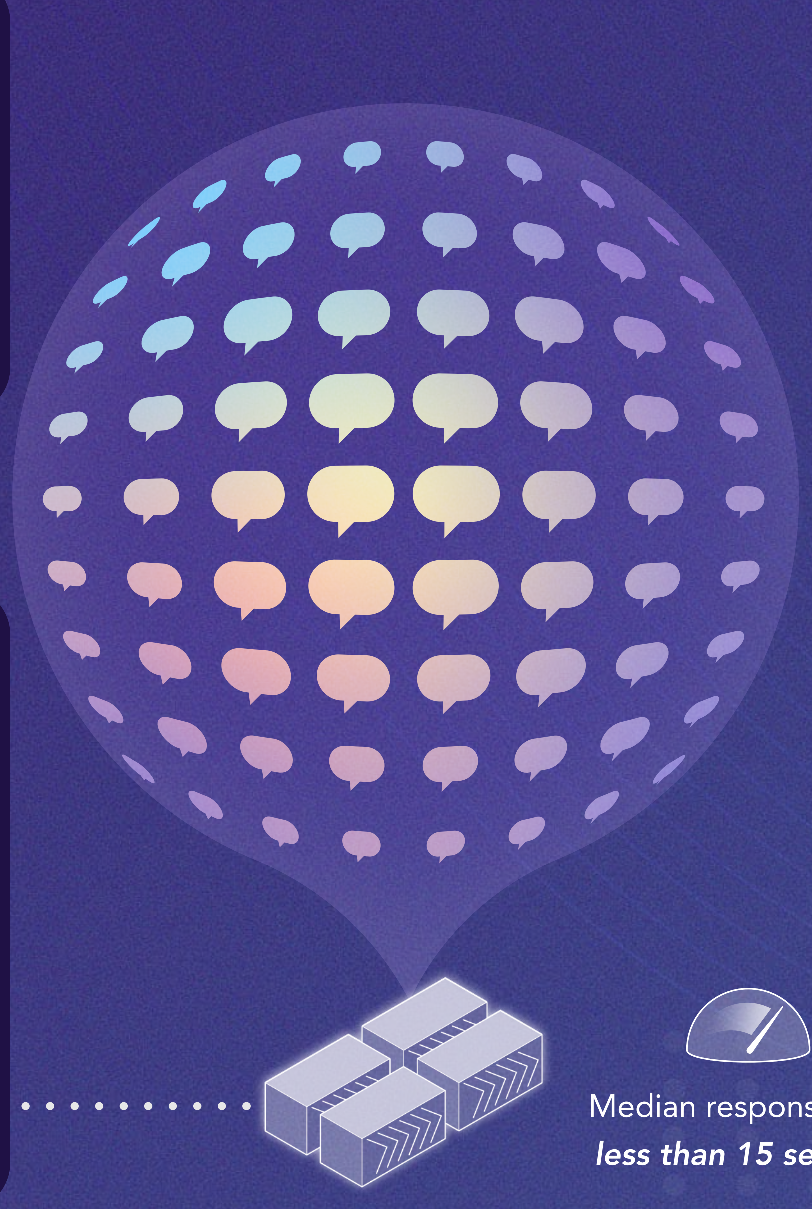
Dell PowerEdge XE9680 servers with AMD Instinct MI300X Accelerators: the power to host GenAI with Llama 3.1 405B LLMs

How can you give your chatbot users faster responses and more meaningful answers to their questions? By taking the data your organization has collected and using it to augment a very large LLM (like Llama 3.1 405B). Our tests proved that running an LLM on the Dell PowerEdge XE9680 powered by AMD Instinct MI300X Accelerators can help make your in-house GenAI project a success.

AMD Instinct™ MI300X Accelerators offer an industry-leading 192 GB of high bandwidth memory (HBM3), so you can...



Support up to

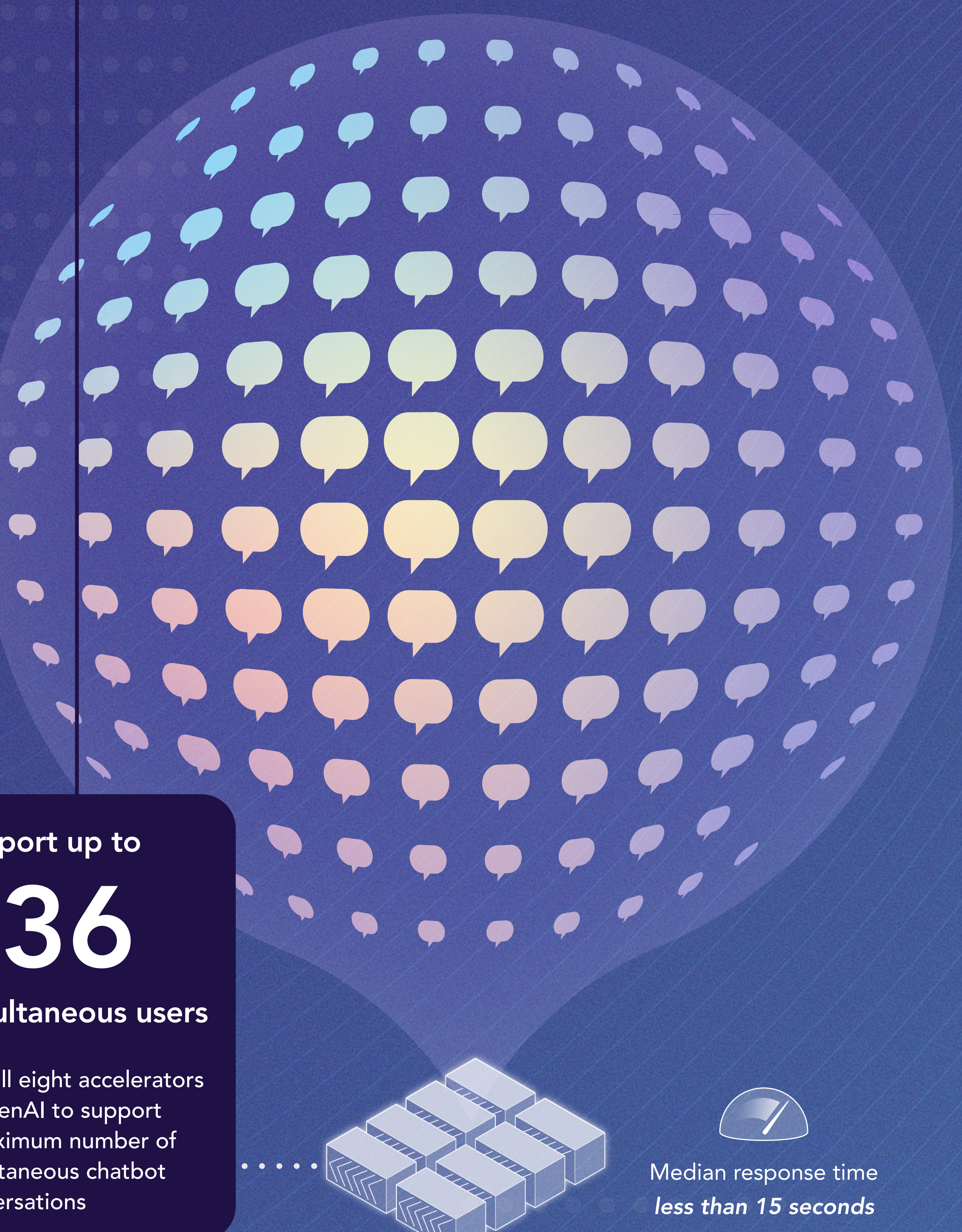
72

simultaneous users

Dedicate half of available accelerators (four) to run your in-house chatbot and use the other four for another application

Median response time *less than 15 seconds*

> *"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[1]



Support up to

136

simultaneous users

Use all eight accelerators for GenAI to support a maximum number of simultaneous chatbot conversations

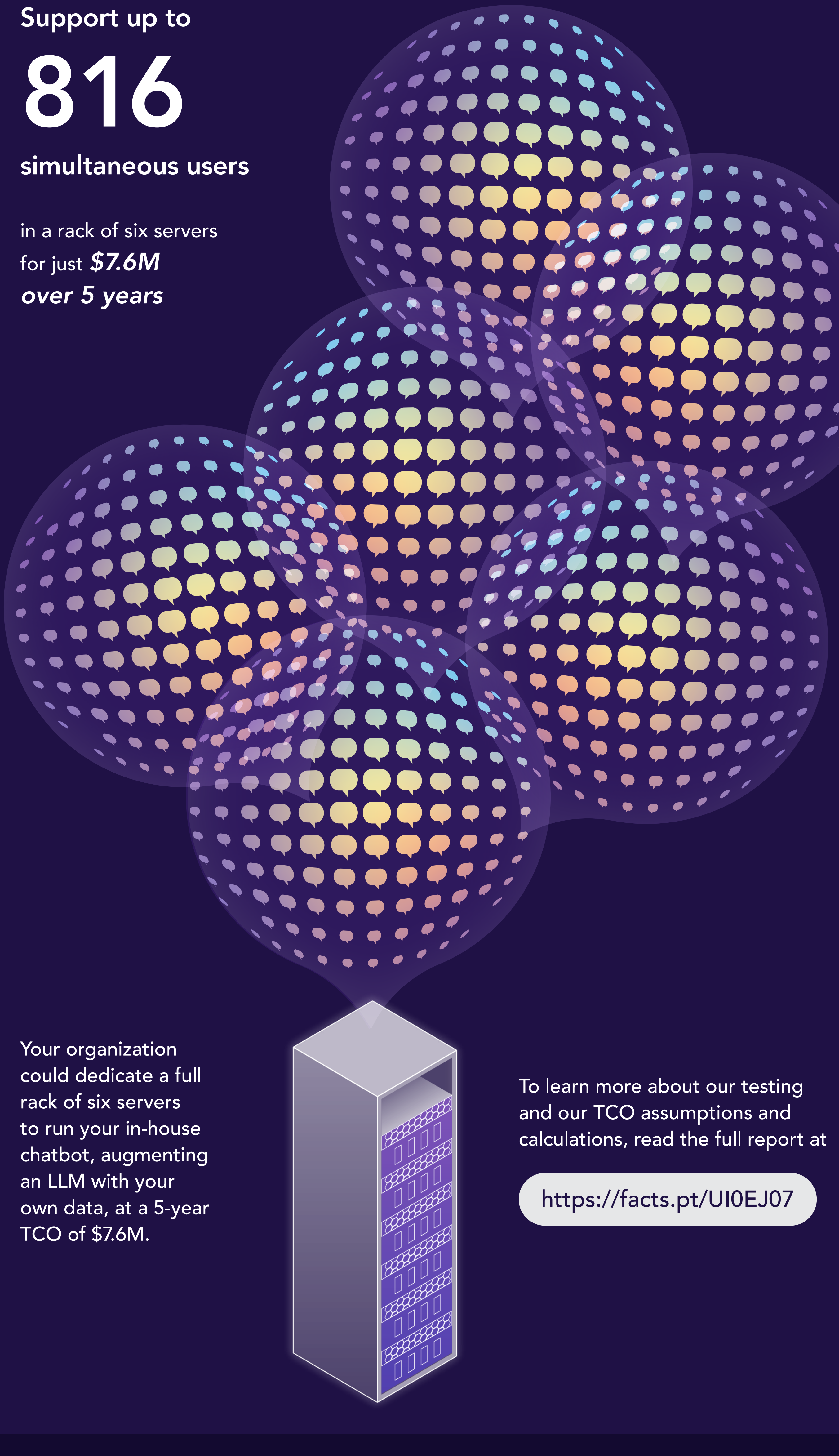
Median response time *less than 15 seconds*

Support up to

816

simultaneous users

in a rack of six servers for just ***$7.6M over 5 years***



Your organization could dedicate a full rack of six servers to run your in-house chatbot, augmenting an LLM with your own data, at a 5-year TCO of $7.6M.

To learn more about our testing and our TCO assumptions and calculations, read the full report at

https://facts.pt/UI0EJ07

[1] 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.

Principled Technologies®

PTChatterly