Enterprise Strategy Group by TechTarget

ECONOMIC SUMMARY

APRIL 2024

Maximizing AI ROI: Inferencing On-premises With Dell Technologies Can Be 75% More Cost-effective Than Public Cloud

Aviv Kaufmann, Practice Director and Principal Validation Analyst

Click to read the full Economic Whitepaper.

Abstract: TechTarget's Enterprise Strategy Group modeled and compared the expected costs to inference large language models (LLMs) on on-premises Dell Technologies infrastructure versus using native public cloud infrastrcture as a service (IaaS) or the OpenAI GPT-4 Turbo LLM model service through an API. We found that Dell Technologies could provide LLM inferencing on premises up to 75% more cost-effectively than native public cloud and up to 88% more cost-effectively than with API services.

Expected Savings When Inferencing LLMs With Dell Technologies Infrastructure

- **38% to 48% more cost-effective than IaaS to inference smaller LLM models (7B parameters)**
- **69% to 75% more cost-effective than IaaS to inference larger LLM models (70B parameters)**
- **Up to 88% more cost-effective than API services to inference larger LLM models (70B parameters)**

Challenges for Enterprises

Organizations are embracing generative AI (GenAI) and LLMs that leverage company-specific data and other intellectual property to automate content generation, answer questions, and make insights readily available to decision-makers. An LLM can be costly and complex to develop from scratch, but organizations can easily augment, fine tune, and customize existing open source LLMs to meet their needs. Organizations can access API-based services such as OpenAI GPT, but inferencing (i.e., querying) costs can quickly add up. Enterprise Strategy Group found that the most popular strategy for organizations to develop and use GenAI supported by an LLM was to utilize an open source LLM and develop a GenAI solution in house.[1] Organizations can build and control their own LLM inferencing solution on powerful GPU-enabled enterprise servers or equivalent GPU-enabled cloud instances and a machine learning platform like NVIDIA's AI Enterprise running open source LLMs like Mistral or Llama.

The Solution – Dell Technologies for LLM Inferencing

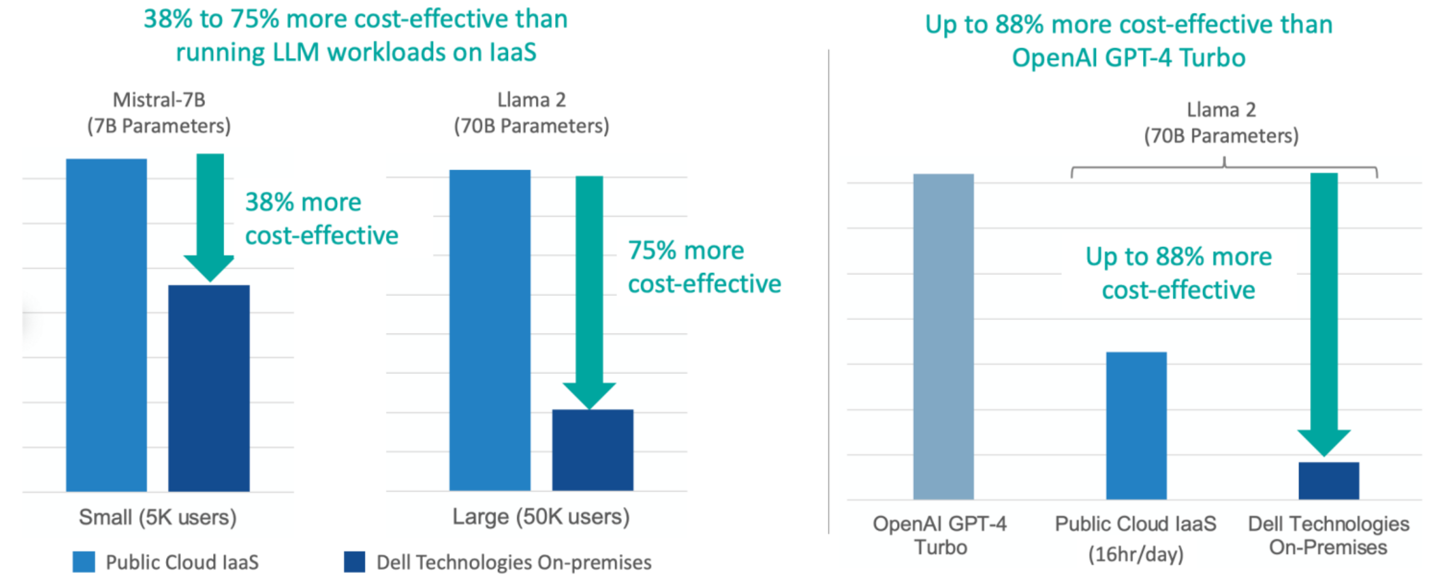
Dell Technologies empowers organizations to bring AI to their data, no matter where it resides, including on-premises edge, colocation, data center, public cloud environments, and on the device itself. Dell simplifies and accelerates organizations' GenAI journeys, creating better outcomes tailored to company needs and safeguarding proprietary data securely and sustainably. In addition to hardware and software infrastructure, Dell offers a robust ecosystem of partners and services to assist organizations, whether they're just starting out or scaling up in their GenAI journey, providing comprehensive solutions that deliver ultimate flexibility now and into the future. To learn more about how Dell can help, visit their GenAI webpage. In addition, with Dell APEX, organizations can subscribe to GenAI solutions and optimize them for multicloud use cases.

[1] Source: Enterprise Strategy Group Research Report, *Beyond the GenAI Hype: Real-world Investments, Use Cases, and Concerns*, August 2023.

Economic Analysis Highlights

Enterprise Strategy Group modeled and compared the expected costs to provide inferencing for 7B- and 70B-parameter text-only LLM models utilizing retrieval-augmented generation (RAG) for organizations of various sizes on Dell Technology hardware, public cloud IaaS on Amazon EC2 instances, and using the OpenAI GPT-4 Turbo API. The Dell Technologies servers and NVIDIA H100 GPU configurations were sized based on the results of inference baseline testing to ensure they would meet usable concurrency and response time parameters. We then considered the associated costs of the hardware, software, support and services, NVIDIA AI Enterprise licenses, power/cooling, and infrastructure and AI platform administration. We sized and priced the native public cloud instances with as-equivalent-as-possible CPU, memory, and H100 GPU capabilities, considering reservation discounts, 16-hour Monday through Friday operations, per-hour NVIDIA AI Enterprise licensing, and cloud administrative advantages. Finally, we modeled the expected costs for users to access the OpenAI GPT-4 Turbo API across various user-generated inference frequencies. The results revealed that Dell Technologies can provide up to 4x more cost-effective inferencing compared to public cloud for these use cases over a three-year period.

Figure 1. Enterprise Strategy Group's Modeled Cost to Handle LLM Inferencing



Source: Enterprise Strategy Group, a division of TechTarget, Inc.

Conclusion

Enterprise Strategy Group strongly recommends that companies implementing LLMs to power their organizations consider taking advantage of the cost-effective technologies and knowledgeable services that Dell Technologies provides to ensure a successful outcome and accelerate their GenAI initiatives.

Click to read the full Economic Whitepaper.

Information contained in this publication has been obtained by sources TechTarget considers to be reliable but is not warranted by TechTarget. This publication may contain opinions of TechTarget, which are subject to change. This publication may include forecasts, projections, and other predictive statements that represent TechTarget's assumptions and expectations in light of currently available information. These forecasts are based on industry trends and involve variables and uncertainties. Consequently, TechTarget makes no warranty as to the accuracy of specific forecasts, projections or predictive statements contained herein.

About Enterprise Strategy Group
TechTarget's Enterprise Strategy Group provides focused and actionable market intelligence, demand-side research, analyst advisory services, GTM strategy guidance, solution validations, and custom content supporting enterprise technology buying and selling.

✉ contact@esg-global.com
www.esg-global.com