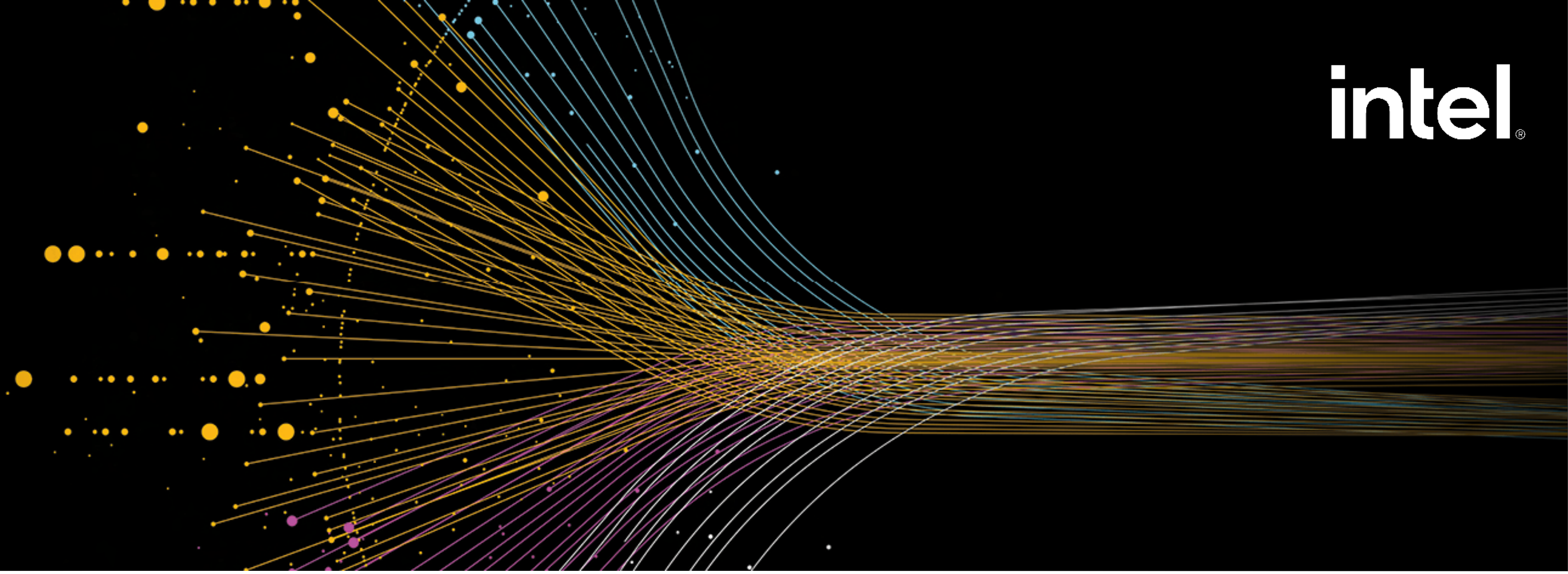

Matching GPU Price Performance Using Amazon Instances With Intel® Xeon® Processors

Among many other services, Storm Reply helps its customers deploy large language models (LLMs) and Generative AI solutions. Storm Reply needed a cost-efficient, high-availability hosting environment to build its LLM-based solution to serve a major company in the energy sector. After a thorough evaluation, a solution developed for the Amazon C7i-family (shared with M7i and R7i) supported by 4th Gen Intel® Xeon® Scalable processors, Intel libraries, and Intel's open GenAI framework proved an ideal hosting environment for Storm Reply's LLM workloads. After optimization, Storm Reply determined that LLM inference on instances with Intel Xeon Scalable processors was on par with GPU instance price performance. Intel libraries also provided a significant benefit. Storm Reply's testing found that the same machine (running Llama 2-13b in bf16 on the same set of questions and same parameters) had an average response time of 92 seconds, contrasting the 485 seconds required without the Intel library.[1]

Products and Solutions
4th Gen Intel® Xeon® Scalable Processors
oneAPI Toolkit
Intel® Extension for Pytorch

Industry
IT Services and IT Consulting

Organization Size
201-500

Country
Italy

Partners
AWS

Learn more
Case Study

[1] For more complete information about performance and benchmark results, visit https://www.intel.com/content/www/us/en/customer-spotlight/stories/storm-reply-customer-story.html