Case Study

Intel® Xeon® Processors

Storm Reply Matches GPU Price Performance Ratio Using Amazon Instances With Intel®Xeon® Scalable Processors

AWS's global footprint, with Intel® AI-ready technologies like custom 4th Gen Intel® Xeon® Scalable processors, Intel® Accelerator Engines, libraries, and GenAI framework, make Large Language Model (LLM) inference easier and more cost-effective.

Solution Summary

- 4th Gen Intel® Xeon® Scalable processors
- Intel® Advanced Vector Extensions 512 (Intel® AVX-512)
- Intel® Advanced Matrix Extensions (Intel® AMX)
- oneAPI Toolkit
- Intel® Extension for Pytorch
- Intel libraries
- Amazon EC2 C7i.16xlarge instances

Executive Summary

Among many other services, Storm Reply helps its customers deploy large language models (LLMs) and Generative AI (GenAI) solutions. Storm Reply chose the new Amazon EC2 C7i instances supported by 4th Gen Intel® Xeon® Scalable processors and Intel libraries for LLM modeling. After a hardware evaluation process, Storm Reply matched the price-performance ratio of GPU-based options by using CPU-based instances. Storm Reply's solution also benefitted from Intel's GenAI framework and the open source LLaMA (Large Language Model Meta AI) model inference using a Retrieval-Augmented Generation (RAG) architecture.

Challenge

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. Storm Reply evaluated several instance types supported by GPUs. However, the GPU-based instances had limitations. First, a shortage of GPUs had the potential to impede Storm Reply's high availability requirements. Secondly, the instances could not customize the RAM allocation on a fixed number of GPU cores, requiring the addition of more GPUs and thus paying a higher price for more RAM. Third, Storm Reply wanted a proven, open solution optimized for LLM inference and GenAI implementation. Lastly, Storm Reply's customer needed the ability to run the trained model locally within his network.

Storm Reply, in partnership with Intel and AWS, brings together technologies and services to deliver customers turn-key, LLM inference and Gen AI solutions to address business priorities.

According to Fortune Business Insights, in today's market with AI surging, more relevant partnerships and collaborations are helping AI-driven companies achieve advancements in artificial intelligence technology.[2]

Storm Reply, in partnership with Intel and AWS, brought together technologies and services to deliver customers turn-key, LLM inference and Gen AI solutions to address business priorities and improve operational efficiencies.

Solution

Regarding AI, some believe GPUs and AI Accelerators are a "must-have." While true for some workloads, a single solution cannot fit all needs. Tailoring the hardware and software hosting the AI Application is an essential step. Thanks to the broad portfolio of solutions, Intel can help find the ideal components to maximize AI application performance developed on AWS platforms.

Storm Reply's strong, long-term relationship with Intel and ongoing knowledge sharing made it easier to discuss challenges and explore possible solutions for Storm Reply's use case. Joint brainstorming and technical enablement activities encouraged Storm Reply to test and consider CPU-based instances. 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.

In addition to the flexibility provided by Intel CPUs, the processors deliver many optimizations and features designed to speed up AI-related workloads. For example, the Intel AVX-512 instruction set improves upon the AVX2 instruction set with wider SIMD registers. The processors also feature Intel® Advanced Matrix Extensions (Intel® AMX), which can offer 3x to 10x higher inference and training performance than the previous CPU generation on bare-metal configuration.[1]

Storm Reply concluded that CPU-based instances offered price-performance similar to GPU environments.[3] Intel-powered solutions added value through easier deployment, excellent scalability, higher instance availability, and flexible RAM allocation. Amazon instances allowed Storm Reply to customize machine resources to optimize workloads and offered greater flexibility in DEV/QA environment dimensions. Storm Reply also made the most of EC2 Spot instances, allowing the company to access "spare" instance resources at a discount during off-peak hours. The instances also proved ideal for Intel's GenAI framework and the open source LLaMA model inference in a retrieval augmented generation (RAG) architecture. To aid in the process, Storm Reply also benefitted from the oneAPI toolkit and the Intel® Extension for Pytorch with the latest performance optimizations for Intel hardware and to accelerate machine learning.

Results

After optimization, Storm Reply determined that LLM inference on instances with Intel Xeon Scalable processors was on par with GPU instance price-performance. On top of this, the ability to customize scripts, combined with Intel's guidance on optimal configurations and proactive assistance, helped Storm Reply implement the solution. 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.

In an era of shared responsibility and sustainability, generative AI and LLMs encourage global change. By working collaboratively, the technology industry provides open, affordable, optimized training and inference solutions on pervasive platforms that give customers greater freedom of choice.

Key Takeaways

- Intel supported Storm Reply's innovations through consultations, technical expertise, and engineering support.
- Storm Reply finds that Intel solutions are ideal AI-ready technology, particularly when hardware availability and configuration flexibility are vital.
- By using CPU-based instances, Storm Reply matched the price-performance ratio of GPU-based options.
- Intel's commitment to innovation, collaboration, and customer needs sets them apart as a reliable and practical choice in AI technology.
- In the age of AI, one solution can't meet all needs. CPUs offer a performant and cost-effective alternative to GPU-based solutions for inference.
- Use Intel libraries to maximize performance.

For More Information

- Learn more about Intel Xeon Scalable processors.
- Get details on Intel Extension for Pytorch.
- See the benefits of the Amazon c7i family.

[1] https://www.intel.com/content/dam/www/central-libraries/us/en/documents/2022-06/enhance-ai-workloads-built-in-accelerators.pdf
[2] https://www.fortunebusinessinsights.com/industry-reports/artificial-intelligence-market-100114
[3] Compared to Amazon EC2 G5 GPU instances.

Performance varies by use, configuration and other factors. Learn more at www.Intel.com/PerformanceIndex.

Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. See backup for configuration details. No product or component can be absolutely secure.

Intel technologies may require enabled hardware, software or service activation.

Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.

No product or component can be absolutely secure.

042024/RJMJ/JS/PDF ♻ Please Recycle