Case Study

Intel® Xeon® Processors

intel ai

Storm Reply Optimizes and Rapidly Deploys Cost-Effective AI Solutions for Infocert and Streamlines DevOps Using CPU Instances

Intel's open GenAI framework and Intel® Xeon® processors provide ideal environment to deploy a large language model (LLM) and Generative AI solution.

Solution Ingredients

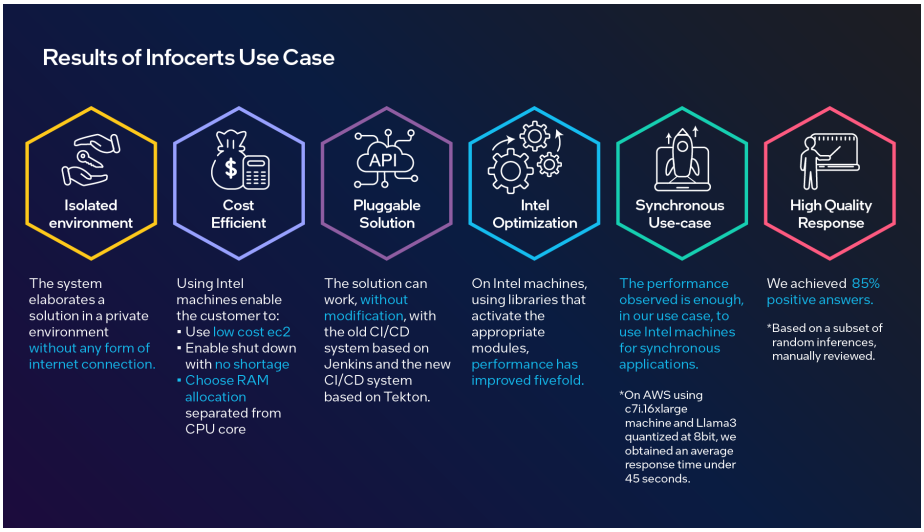
- Amazon EC2 Instances
- 4th Gen Intel® Xeon® processors
- Intel® Advanced Matrix Extensions (Intel AMX)
- Intel® Advanced Vector Extensions 512 (Intel® AVX-512)
- Intel® Extension for Pytorch

Executive Summary

As a systems integrator, Storm Reply helps its global customers plan and implement cloud-based solutions and services. Amazon Web Services and Intel hardware and software technologies provide Storm Reply with the building blocks to bring the power of AI solutions into customers' business environments. For example, Storm Reply helped Infocert deploy a large language model (LLM) and Generative AI solution with a strategy that provided economic and performance benefits. Storm Reply elected a variety of Amazon EC2 instances supported by 4th Gen Intel® Xeon® Scalable processors, Intel® libraries, and Intel's open GenAI framework as an ideal environment to host the solution. Storm Reply also used Intel technologies to enable a Retrieval Augmented Generation (RAG) solution that helped Infocert automate tasks and streamline the DevOps process by identifying and fixing pipeline errors.

Challenge

Storm Reply needed a cost-efficient, high-availability hosting environment to build its LLM-based solution for Infocert. Storm Reply evaluated several instance types



Storm Reply helped Infocert deploy a large language model (LLM) and Generative AI solution based on Intel® technologies with a strategy that provided economic and performance benefits.

supported by GPUs. However, those environments had three limitations:

- A GPU shortage could hamper Storm Reply's high availability requirements.
- GPU-based instances could not customize the amount of RAM allocated per GPU core. This limitation made adding more RAM necessary to optimize GPU-driven workloads, which added cost compared to instances supported by Intel Xeon processors.
- Storm Reply's customer needed a reliable and open solution tailored for LLM inference and GenAI deployment and a trained model operable within a local network.
- The company also faced the challenge of identifying pipeline errors rapidly and accelerating fixes to correct them to reduce ops costs and time to market.

Solution

Storm Reply's solution developed for the Amazon EC2 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 the LLM. The CPUs deliver many optimizations and features designed to accelerate AI-related workloads. For example, the Intel® Advanced Vector Extensions 512 (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 also benefitted from the oneAPI toolkit and the Intel® Extension for Pytorch with the latest performance optimizations for Intel hardware to accelerate machine learning.

Storm Reply also has a unique solution, supported by Intel and Amazon technologies, for customers seeking to find and correct errors in the task automation pipeline. The company's Integrated DevOps Intelligent Assistant (IDIA) analyzes continuous integration (CI) and continuous delivery (CD) pipeline errors using a specialized LLM. When an agent finds errors, it also identifies the related source code. The issue then moves to Storm Reply's RAG system. The solution helps to identify the most probable culprits and give the GenAI engine proper context about the error. The RAG output then helps create a prompt to instruct the engine for remediation. Once prepared, the prompt moves to the LLM engine hosted on Amazon EC2 instances and returns a resolution.

> "With the help of Amazon and Intel technologies, we can help our customers deploy AI rapidly and streamline DevOps. By simplifying and accelerating these processes, customers like Infocert can save time and gain rapid ROI."
>
> *—Alessandro De Carolis, Business Unit Manager, Storm Reply*

Global	
Totally Useless	2%
Something useful but largely useless	3%
Neutral	11%
Mostly useful, with some imprecision	38%
Almost entirely correct and useful	47%

A joint test on a subset of LLM responses provided 85% "useful" or "very useful" answers. The average response time from the Intel machine is 49 seconds, which drops to 42.8 seconds when removing a small percentage of outliers.

Results

Storm Reply concluded that a solution for Infocert using CPU-based instances offered price performance similar to GPU environments.[2] However, Intel-powered solutions added value through easier deployment, excellent scalability, higher instance availability, and flexible RAM allocation. Amazon EC2 instances allowed 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 RAG architecture. Intel libraries provided a significant benefit through reduced latency. 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 with the 485 seconds required without the Intel library.[3] With further enhancement on AWS using the same EC2 C7i.16xlarge machine and Llama 3 quantized at 8bit, Storm Reply reduced the average response time to under 43 seconds[4] while maintaining the same response quality.

For Storm Reply's customers who need help streamlining DevOps work for task automation, Intel and Amazon technologies also play a vital role. By identifying problems and resolving them quickly, Storm Reply's customers using its IDIA system can shave significant time off the process.

The solution also offers Infocert additional benefits:

- Intel optimization: Using libraries that activate the appropriate modules, performance improved nearly fivefold on Intel machines.[3]
- Isolated Environment: The system can work in a private environment without an internet connection.
- Cost efficient: Amazon EC2 instances with Intel CPUs are cost-effective, enabling shutdown with no shortage, and choosing RAM allocation for CPU cores.
- Pluggable solution: The solution can run on the legacy Jenkins-based CI/CD system and the new CI/CD system on Tekton without modification.

Key Takeaways

- Through Intel solutions, Storm Reply's portfolio of technologies can deploy AI anywhere from the Amazon cloud to the edge.
- Storm Reply taps GenAI to help customers identify pipeline errors and solve them.
- Intel offers optimized hardware with built-in AI acceleration, software, and toolkits to speed up AI usage scenarios.
- Open standards for building AI applications create value for partners and developers alike.

intel + aws

[1] https://www.intel.com/content/dam/www/central-libraries/us/en/documents/2022-06/enhance-ai-workloads-built-in-accelerators.pdf
[2] Compared to Amazon EC2 G5 GPU instances.
[3] Intel libraries allow for maximum utilization of the latest-generation Intel CPUs.
[4] On AWS using EC2 C7i.16xlarge machine and Llama 3 quantized at 8bit, Storm Reply obtained an average response time of under 43 seconds.
[5] The Storm Reply LLM/RAG solution provided Inforcert with 85% "useful" or "very useful" responses based on a subset of random inferences, manually reviewed.

112024/RJMJ/DO/PDF ♻ Please Recycle 363110-001US