Case Study

Intel® Xeon® Processors

intel ai

Aible Serverless AI Helped Fortune 500 Company Implement and Gain Value in Eight Days

Amazon Lambda and Intel® Xeon® processors enable Aible's AI deployments quickly and cost-effectively with a pay-as-you-go model.

Solution Ingredients

- Amazon Lambda
- Intel® Xeon® processors
- Intel® Advanced Vector Extensions 512 (Intel® AVX-512)
- Intel® Vector Neural Network Instructions (Intel® VNNI)

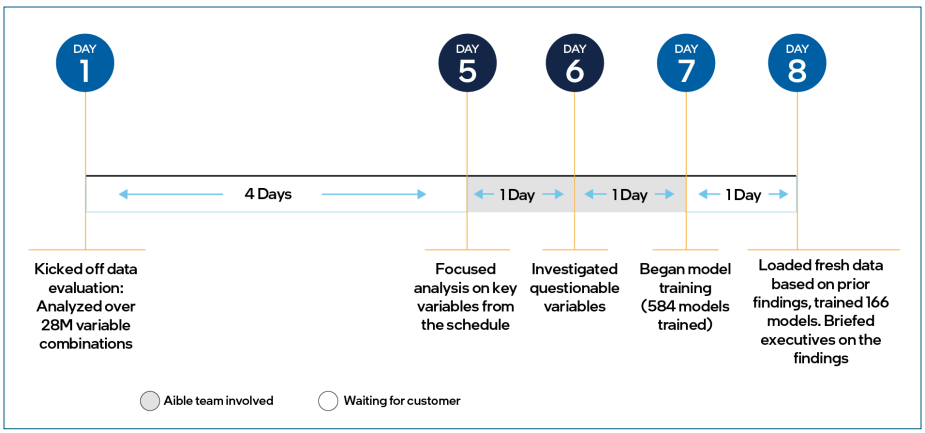
Executive Summary

Aible's mission is to help enterprises embrace AI solutions easily, quickly, and cost-effectively. Aible's solution takes a unique "serverless" approach running on Amazon Lambda and supported by Intel® Xeon® processors. Rather than running a Generative AI (GenAI) query on a virtual server, for example, the request can run directly on Amazon Lambda-hosted CPUs, using only the compute resources necessary to handle that specific task. Because customers pay only for the actual time using processors, Aible AI solutions prove highly cost-effective. A recent benchmarking study by Aible and Intel found that customers deployed and derived real value from AI in as little as eight days[1] and could demonstrate a 55X reduction in total cost of ownership for GenAI workloads.[2]

Challenge

Implementing AI in the enterprise can prove challenging. Customers typically require in-house expertise and a significant investment in on-premises or cloud infrastructure. Deploying AI and GenAI models using traditional methods can involve several months of planning, model training, testing time, and adequate cloud instance bandwidth to handle AI workloads when required.

Customers also need a turnkey way to deploy GenAI while maintaining the security protocols and governance of their data sets while minimizing the cost of cloud



Aibles' very rapid AI deployment enabled one transportation industry enterprise to deliver an AI solution in only eight days.

Aible's serverless approach supported by Intel® Xeon® processors could demonstrate up to a 55X reduction in total cost of ownership for GenAI workloads.

resources through a pay-as-needed model. Time-to-value is also a critical factor for customers. Some spend millions of dollars making GenAI investments and need the fastest ROI possible.

Solution

Amazon Lambda with Intel Xeon processors allows customers to run workloads without managing or provisioning servers or optimizing their applications for specific instance types and hardware. By building its technologies to run on Amazon Lambda, Aible innovated a new approach to AI implementation using a serverless architecture.

Lambda allocates compute resources only when user requests come in and can scale up or down dynamically as needed. Users pay only for the processing time their Gen AI inquiry consumes. AI acceleration technologies built into Intel Xeon processors, like the Intel® AVX-512 accelerator and Intel® VNNI, can handle AI-based tasks without needing GPUs.

Because Aible-enabled AI runs serverless, the Amazon Lambda-based solution proves extremely cost-effective since different customers can share CPU resources securely and pay for only the processing resources needed.

Results

Aible enables very rapid GenAI deployment. As illustrated by the timeline below, one transportation industry enterprise had an AI solution in place after only eight days.[3]

In the past, moving an AI solution from the planning to the deployment stage could take several months. In contrast, Aible's functionality allows customers to derive almost immediate value.[1] For example:

- A Fortune 500 Healthcare Provider found new insights in Social Determinants of Health (SDoH) data with a 20X improvement in speed to insight in 15 days.
- Nova Southeastern University used Aible solutions to potentially improve student retention by 17% in 15 days.
- A multinational company selling beauty and cosmetics products used AI to identify ways to drive $10M in additional sales by optimizing first orders in 17 days.
- A global food company identified ways to reduce food wastage by more than 10% in 27 days.
- A global manufacturer identified ways to reduce the impact of late shipments by more than $4M annually in just 17 days.

During a testing and benchmarking process comparing several generations of Intel Xeon processors, Aible found the cost for serverless AI usage can generate up to a 55x cost reduction in total cost of ownership for GenAI workloads.[2]

> "Because of Aible's serverless approach to AI deployment using Amazon Lambda and Intel Xeon processors, we can allow many users to share resources with others securely, and they can tackle AI tasks much more cost-efficiently. Our solutions can demonstrate a 55X reduction in total cost of ownership for Generative AI workloads. The combination of cost reduction and the speed of implementing the entire use case securely in the customer's cloud in days truly makes enterprise GenAI accessible to everyone today."
>
> *—Arijit Sengupta, Founder and CEO, Aible*

Resources

- Learn about Aible serverless solutions for AI.
- Read about Amazon Lambda and its benefits.
- Find out more about Intel Xeon processors.

[1] https://enaible.aible.com/hubfs/Intel_Aible_Benchmark.pdf
[2] https://www.youtube.com/watch?v=07pOMeeb3BA&t=121s
[3] https://enaible.aible.com/hubfs/Brochure_Aible_Intel_Global_Airline.pdf

102024/RJMJ/JS/PDF ♻ Please Recycle 362791-001US