Case Study

Intel® AI Solutions
Intel® Xeon® Scalable Processors

Aible Delivers Generative AI at up to 55x Lower Cost

Aible clients get AI into production fast, consistently see business value in under 30 days.

At a glance

- 5th Gen Intel® Xeon® Scalable processors increase Generative AI (GenAI) performance 11%[1] over 4th Gen Intel® Xeon® Scalable processors
- Aible delivers GenAI with RAG services on standard, serverless Intel® CPU cloud instances—no expensive GPUs or accelerators required
- According to Aible's benchmark analysis, customers can realize up to a 55x cost savings when running RAG models on their CPU-based serverless solutions[2]
- Customers typically cut months off their development time with Aible's end-to-end AI platform

Executive summary

Aible clients consistently get AI projects up and delivering business value in under 30 days. In tests, Intel and Aible demonstrated that the most common Generative AI (GenAI) workload—GenAI with RAG—performed well on serverless CPUs at projected costs up to 55x lower than standard, server-based AI services.[1]

Challenge

According to a 2024 Gartner® press release, "On average, only 48% of AI projects make it into production, and it takes 8 months to go from AI prototype to production." [3] Generative AI (Gen AI) may be making headlines, but businesses don't seem to be benefiting.

Solution

The Aible AI platform combines user-friendly tools for GenAI with low-cost serverless, CPU computing infrastructure. By simplifying GenAI toolsets and lowering costs, Aible makes it possible for developers to iterate quickly, and for business users to achieve business alignment and build AI applications on their own.

Results

Intel tested the most common GenAI) workload—GenAI with RAG—using the Aible platform on three generations of Intel® Xeon® Scalable processors. Tests show widely available 2nd Generation Intel® Xeon® Scalable instances are fully capable of running Gen AI with RAG—without GPUs or other accelerators. Additional tests demonstrated GenAI responses in seconds with up to 20% faster performance on 4th Generation[1] and up to 30% faster performance on 5th Generation Intel® Xeon® Scalable processors.[1]

According to Aible's benchmark analysis, running Aible Gen AI with RAG on serverless CPUs cuts ongoing computing costs up to 55x.[2]

Aible serverless platform proves CPUs solve key development barriers for Gen AI

"When I looked at why AI projects fail, it's always because of disconnect between the data scientists, experts, and business users," says Aible founder and CEO Arijit Sengupta. "So, what we obsessively focus on is, how can I get the business user in the mix as quickly as possible?"

The high cost of AI computing creates scarcity that makes it prohibitively expensive for AI developers to iterate at scale and fail fast. Tight budgets can squeeze subject matter experts out of the process. Aible tackles computing costs with a serverless approach that delivers GPU-level performance by scaling across multiple CPUs to create a distributed, parallel-processing instance. By using serverless CPU instances, the Aible platform lowers computing costs so that data scientists, subject matter experts, and business owners can be involved throughout the AI development process.

It's an excellent model, but can serverless CPUs meet the voracious demands of Gen AI? To find out, Intel and Aible benchmarked one of the most common GenAI workloads—GenAI with RAG—with the Aible platform running on 2nd, 4th, and 5th Generation Intel® Xeon® Scalable processors.

The results were very encouraging. On a single 5th Generation Intel® Xeon® Scalable processor, GenAI with RAG execution times were as low as 21.02 seconds.[1] Testing showed a single, 2019-era 2nd Generation Intel® Xeon® Scalable processor finished in 30.2 seconds.[1]

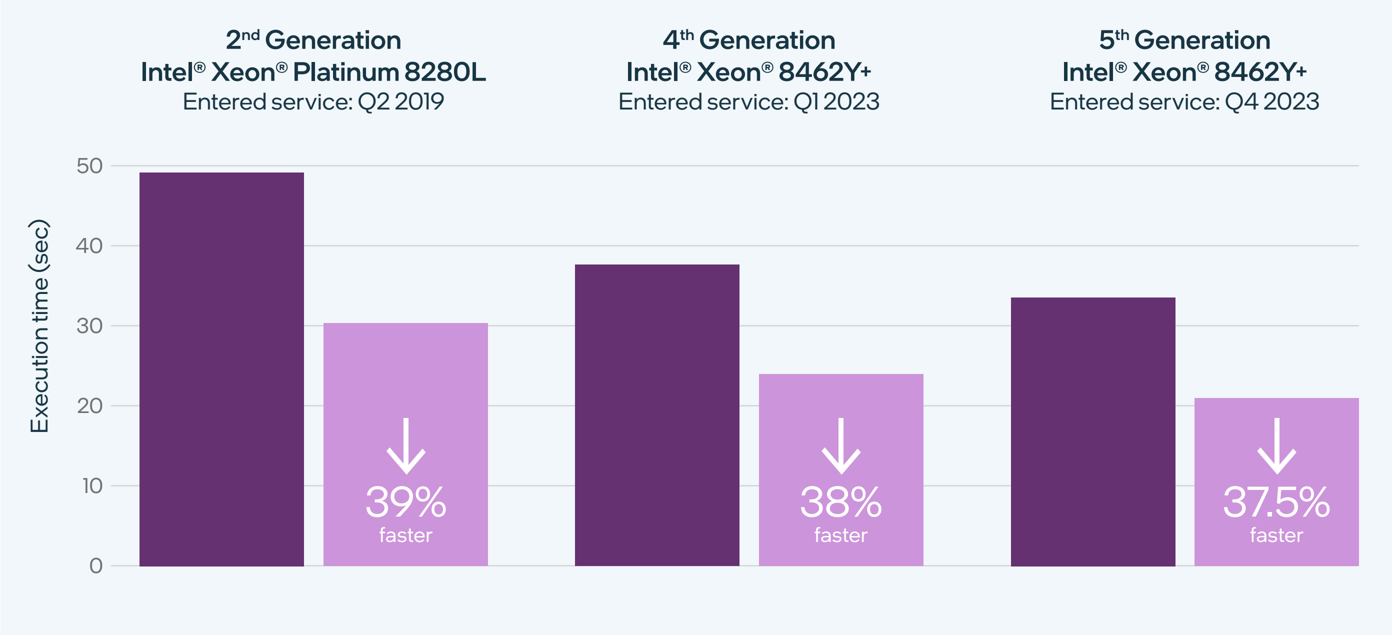
Since the Aible serverless approach scales across multiple CPUs, Aible platform customers can expect highly responsive GenAI services from standard Intel® Xeon® instances.

Serverless compute is up to 55x[2] less expensive for Gen AI

To compare real-world costs, Aible priced AI with RAG services on three computing configurations.

- Shared public cloud LLM with a leading hosted VectorDB
- LLM and VectorDB running on dedicated servers in a private cloud
- LLM and VectorDB on the Aible platform using cloud-based, serverless CPU instances

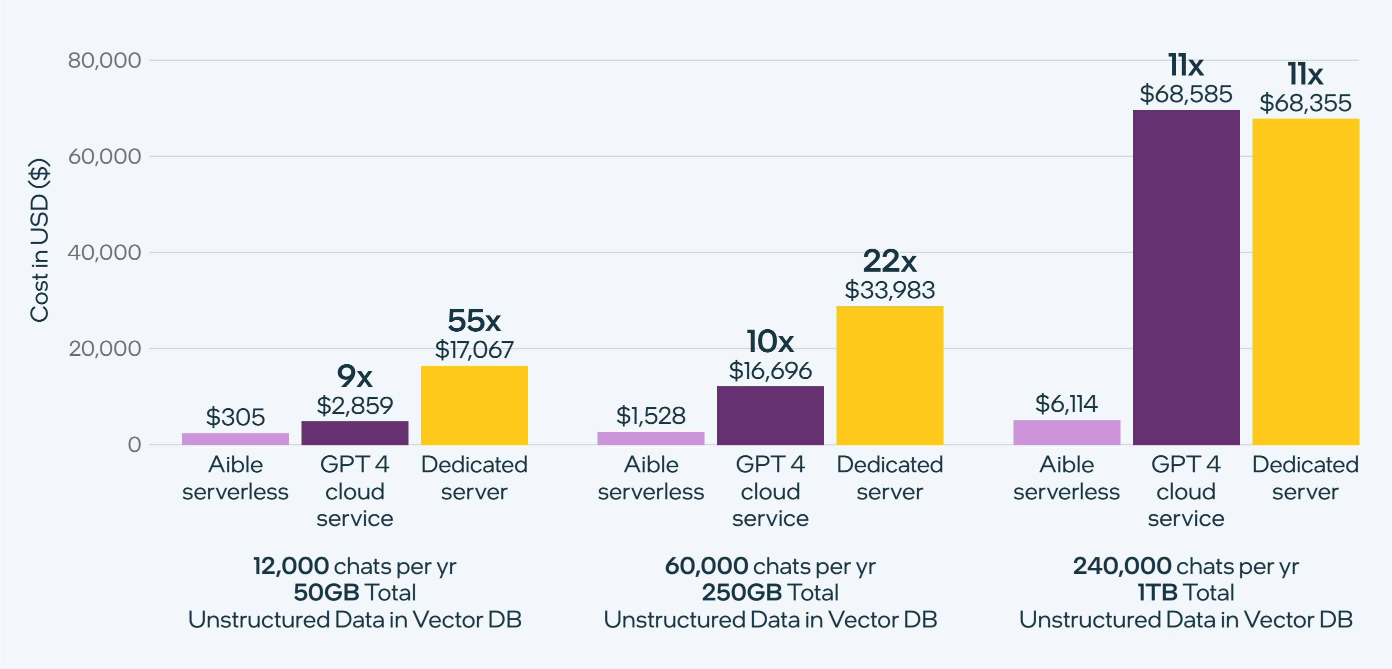
Aible benchmark results on Intel CPU without GPU (lower is better)[1]



Why did Intel and Aible test GenAI with RAG on six-year-old CPUs?

Most serverless instances only expose "lowest common denominator" technologies, which are roughly equivalent to a ten-year-old CPU. CPUs from 2019 provide a more realistic benchmark.

Estimated annual costs fro GenAI with RAG (lower is better) [2]



Aible then projected costs for small (12,000 chats/50GB of unstructured data), medium (60,000 chats/250GB of unstructured data), and large GenAI (240,000 chats/1TB of unstructured data) AI with RAG workloads on each configuration. In each case, serverless computing infrastructure was significantly less expensive—from 9x up to 55x less.[2]

How Aible exploits serverless computing quirks to increase AI performance

True serverless computing provides processing power on demand. The user doesn't have to reserve or pay for a server in preparation for a workload. This means nothing happens—and no costs are incurred—until a user asks a question.

One downfall of serverless is the time it takes to spin up. Aible solves this by taking advantage of the fact that serverless instances persist after they've completed a task. By carefully queueing their platform's active workloads, Aible keeps serverless instances running non-stop—delivering near-dedicated server performance with no time lost to idling down and spinning back up.

Dynamic model selection reduces costs even further

Aible provides proactive control over the ongoing costs of AI services. Administrators can budget costs per chat and costs per user to limit overspending. The system can dynamically change models to match costs to budgets. For example, the system can switch individual users and entire use cases from an expensive, general-purpose LLM to a smaller, less expensive model like Mistral or Llama 8B.

Lower cost and friendlier tools create new techniques for getting AI into production

The Aible platform's ease of use and low-cost computing support faster, scrappier approaches to AI development. With Aible, clients can fail fast and get into production without breaking the bank.

- Run 10, 50, or 100 model variations simultaneously
- Test RAG, fine-tuning, and few-shot learning, all at the same time
- Test and refine multiple prototypes with business users
- Bring Gen AI augmentation to data engineering, analytics, and machine learning

Because they can work faster, Aible clients typically see business value in less than thirty days, some in as few as eight days.[4]

Intel and Aible – A partnership that is bringing AI everywhere

Aible customers are proof that lowering computing costs and simplifying AI development fundamentally change how AI projects move through development and into production. A leading analyst firm states that it can take over 8 months to complete a generative AI project. "We have customers who have completed their projects in just two days," says Aible CEO Arijit Sengupta.

Making AI more accessible and successful isn't just about fast paths to business value. It's about unlocking AI's potential and expanding its application for all. "Our motto from the beginning was I-am-able because we believed anyone should be able to create AI that serves their needs," says Sengupta. "We are solving that fundamental risk point of AI—that most of the projects are going to fail. If we change the way AI is done—move it away from monolithic AI to AI that empowers the individual—we will have succeeded.

About Aible

Aible provides user-friendly tools at lower costs so everyone can develop their own AI and have a say in how it affects their lives.

Learn more at aible.com

About Intel

Intel efforts in AI include hardware and performance-improving AI software design and optimization. Intel works with partners of every size to bring AI everywhere.

Learn about Intel AI Tools

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

Gen AI with RAG configuration details:

1-node, 2x Intel® Xeon® Platinum 8280L CPU @ 2.70GHz, 28 cores, HT On, Turbo On, NUMA 2, Integrated Accelerators Available [used]: DLB 0 [0], DSA 0 [0], IAA 0 [0], QAT 0 [0], Total Memory 384GB (12x32GB DDR4 2933 MT/s [2934 MT/s]), BIOS SE5C620.86B.02.01.0017.110620230543, microcode 0x5003604, 2x Ethernet Connection X722 for 10GBASE-T, 1x 894.3G INTEL SSDSC2KB96, 1x 1.8T INTEL SSDPE2KX020T8, 2x 3.7T INTEL SSDPE2KX040T8, Red Hat Enterprise Linux 8.9 (Ootpa), 4.18.0-513.18.1.el8_9.x86_64, WORKLOAD=Aible End-to-end RAG-LLM, Model=Mistral-7B-OpenOrca-GGUF, all-MiniLM-L6-v2, gcc 12.2.0, IntelLLVM 2024.0.2, llama.cpp, ChromaDB, Langchain, oneAPI base container 2024.0.1-devel-ubuntu22.04. Tested by Intel on 03/07/24.

1-node, 2x Intel® Xeon® Platinum 8462Y+, 32 cores, HT On, Turbo On, NUMA 2, Integrated Accelerators Available [used]: DLB 2 [0], DSA 2 [0], IAA 2 [0], QAT 2 [0], Total Memory 512GB (16x32GB DDR5 4800 MT/s [4800 MT/s]), BIOS 05.12.00, microcode 0x2b0004d0, 2x BCM57416 NetXtreme-E Dual-Media 10G RDMA Ethernet Controller, 2x Ethernet Controller E810-C for QSFP, 2x 3.5T SAMSUNG MZQL23T8HCLS-00B7C, 1x 1.8T SAMSUNG MZ1L21T9HCLS-00A07, Red Hat Enterprise Linux 8.9 (Ootpa), 4.18.0-513.18.1.el8_9.x86_64, WORKLOAD=Aible End-to-end RAG-LLM, Model=Mistral-7B-OpenOrca-GGUF, all-MiniLM-L6-v2, gcc 12.2.0, IntelLLVM 2024.0.2, llama.cpp, ChromaDB, Langchain, oneAPI base container 2024.0.1-devel-ubuntu22.05. Tested by Intel on 03/07/24.

1-node, 2x INTEL(R) XEON(R) PLATINUM 8562Y+, 32 cores, HT On, Turbo On, NUMA 2, Integrated Accelerators Available [used]: DLB 2 [0], DSA 2 [0], IAA 2 [0], QAT 2 [0], Total Memory 512GB (16x32GB DDR5 5600 MT/s [5600 MT/s]), BIOS 3B05.TEL4P1, microcode 0x21000161, 2x Ethernet Controller X710 for 10GBASE-T, 2x Ethernet Controller E810-C for QSFP, 1x 894.3G INTEL SSDSC2KG96, 1x 3.5T SAMSUNG MZQL23T8HCLS-00A07, 3x 3.5T SAMSUNG MZQL23T8HCLS-00B7C, Red Hat Enterprise Linux 8.9 (Ootpa), 4.18.0-513.18.1.el8_9.x86_64, WORKLOAD=Aible End-to-end RAG-LLM, Model=Mistral-7B-OpenOrca-GGUF, all-MiniLM-L6-v2, gcc 12.2.0, IntelLLVM 2024.0.2, llama.cpp, ChromaDB, Langchain, oneAPI base container 2024.0.1-devel-ubuntu22.06. Tested by Intel on 03/07/24.

[2] Aible, Benchmark Report, Serverless Approach for Generative AI on CPUs, accessed August 2024. Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.

[3] Gartner, Gartner Survey Finds Generative AI Is Now the Most Frequently Deployed AI Solution in Organizations, May 7, 2024.

[4] Intel and Aible, Case Studies from the "Impact from AI in 30 Days" Program, accessed July 2024. Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.

Gartner Disclaimers

Notices & Disclaimers

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

Performance varies by use, configuration, and other factors. Learn more on the Performance Index site.

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.

Your costs and results may vary.

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

1124/SB/CAT/PDF ♻ Please Recycle 363130-001EN