Brief

Cloud Services
Intel® Gaudi® 3 AI Accelerator

Helping Enterprises Scale AI with Intel® Gaudi® 3 AI Accelerators on IBM Cloud

Gain affordability, performance, and scalability along with resilience, security, and compliance—all with easily deployable solutions in the hybrid cloud.

At a glance

IBM Cloud is the first cloud services provider to make Intel® Gaudi® 3 AI accelerators available to its enterprise customers.[1] Intel Gaudi 3 is designed for:

- Cost-effectiveness, security and resiliency
- Flexibility with an open ecosystem
- Hosting bigger models using fewer accelerators due to large memory[2]

Executive summary

From the inception in 1981 of the first IBM personal PC, powered by the 8088 microprocessor, Intel and IBM have been helping customers and partners consistently achieve their goals. This long-term collaboration has driven innovation and delivered exceptional results across various technological eras. Today, IBM and Intel are continuing that collaborative success by deploying Intel® Gaudi® 3 AI accelerators as a service on IBM Cloud. This offering aims to help enterprises more cost-effectively scale AI and drive innovation while prioritizing openness, security, and resiliency. This collaboration will also enable support for Intel Gaudi 3 AI accelerators within IBM's watsonx AI and data platform in Q2 2025.

Challenge

There's no doubt that AI can help drive enterprise efficiency and competitiveness. For example, generative AI (GenAI) can boost HR productivity by up to 8x,[3] handle 97% of customer questions,[4] and generate 60% of software development content.[5]

However, enterprises struggle to reach such tantalizing goals due to several cost and scalability concerns; the time and effort involved in choosing and deploying infrastructure that meets performance goals; and security, compliance/governance, and resilience requirements. These same challenges are compounded when enterprises consider deploying AI workloads in the hybrid cloud. Enterprises need trusted partners that can help them unlock the potential of GenAI.

Solution

IBM and Intel have a long-standing collaborative relationship, and each have a vision of creating AI systems with total cost of ownership (TCO) advantages and enabling an open ecosystem to scale enterprise AI.

IBM Cloud is the first cloud service provider to make Intel® Gaudi® 3 AI accelerators available to its enterprise customers.[1] IBM Cloud offerings include Intel Gaudi 3 AI accelerators as part of its full-stack approach to GenAI.

Intel Gaudi 3 accelerators include AI-specific design and AI-centric features: These devices are specifically designed to help meet the exploding demands for GenAI, large model inferencing, and model fine-tuning while supporting an open development framework. In particular, Intel Gaudi 3 accelerators are ideal for multi-model large language models (LLMs) and retrieval-augmented generation (RAG), which helps streamline the integration with IBM's watsonx and data platform. Intel Gaudi 3 accelerators feature matrix math engines, tensor processing cores, high-bandwidth memory, and built-in Ethernet ports for accelerated inferencing of deep neural networks.

> "Leveraging Intel Gaudi 3 accelerators on IBM Cloud will provide our clients access to a flexible enterprise AI solution that aims to optimize cost performance. We are unlocking potential new AI business opportunities, designed for clients to more cost-effectively test, innovate and deploy AI inferencing solutions."
>
> — Alan Peacock, general manager, IBM Cloud

Expected business outcomes

- **Competitive AI performance:** Intel Gaudi 3 accelerators offer significant performance-related advancements over Intel® Gaudi® 2 accelerators, with 4x the compute, 2x the networking bandwidth, and 1.5x the memory bandwidth. They also have 128 GB of high-bandwidth memory (HBM) capacity at 3.7 TB/sec bandwidth to speed up GenAI performance.[2]

High performance

Low cost

Open and flexible

1. Intel® Gaudi® 3 AI accelerator benefits.

- **Cost-conscious design:** IBM testing showed that on a single card, Intel Gaudi 3 AI accelerators on IBM Cloud can generate over 5,000 tokens per second for IBM's granite-8b model, supporting over 100 concurrent users with an inter-token latency of less than 20 milliseconds. This equates to ~50 average length emails.[6]
- **Scale-up and scale-out potential:** Enterprises can scale from a single node (eight accelerators) with a throughput of 9.6 TB/s to a 1,024-node cluster (8,192 accelerators) with a throughput of 9.830 PB/s. Scaling is achieved using a choice of numerous industry-standard and high-capacity Ethernet switches and other supporting infrastructure to help lower costs.[2]

  As stated above, IBM testing shows a linear performance scale factor of over 5,000 tokens per second for IBM's granite-8b model.[6] Enterprises need this kind of scalability to efficiently keep up with the growing demand for AI compute.
- **Support for open development:** IBM has been a driving force in the evolution of open-source technology for over twenty years. From the hypervisor to developer tools to blockchain and AI, IBM Cloud is built on open-source technologies like Linux and Kubernetes. IBM specializes in operationalizing, hardening, scaling, and contributing back to the open-source ecosystem.

  Intel Gaudi 3 accelerators are designed for broad AI application support. They are fully compatible with the open PyTorch framework (used for nearly all GenAI development), which helps enable easy integration and migration. Developers can easily leverage Intel Gaudi 3 AI accelerators' advanced capabilities for AI innovation to help reduce development time and code maintenance.

Intel Gaudi 3 accelerators are helping IBM achieve its continued efforts to address the challenges enterprises face in achieving the required compute power for GenAI workloads, balanced with a cost-conscious entry point. Powering AI workloads with Intel Gaudi 3 accelerators aims to help lower TCO while enhancing performance.

By running their GenAI workloads on Intel Gaudi 3 accelerators, enterprises can access new AI business opportunities, including those in highly regulated industries, to more cost effectively test, innovate, and deploy AI inferencing solutions, which drives the ability to scale enterprise AI with optimized price/performance. Additionally, integrating Intel Gaudi 3 AI accelerators into IBM Cloud Virtual Servers for Virtual Private Cloud (VPC) aims to help enable x86-based enterprises to run applications fast and securely, enhancing user experiences.

A closer look at Intel Gaudi 3 AI accelerators and IBM Cloud synergy

Common enterprise GenAI use cases include chatbots and virtual assistants, code generation, natural language translations, and text summarization and paraphrasing. As discussed in the previous section, Intel Gaudi 3 accelerators are ideal for these GenAI workloads. However, the advantages of using IBM go far beyond choosing a specific hardware offering.

While enterprises will benefit from Intel Gaudi 3 accelerators, they also benefit from IBM's long experience in providing an enterprise cloud platform that helps meet the needs of heavily regulated sectors, such as financial services, government, healthcare, and telco. Other industries, such as retail and media, can also leverage IBM's hybrid cloud by design strategy.

Designed for security and AI governance

Security is paramount when it comes to AI and sensitive enterprise data. Enterprises need to be assured of the security of their data so that they can enjoy the benefits of GenAI without exposing their proprietary data and ideas to the open Internet.

To meet this concern, AI must be deployed responsibly, with end-to-end AI lifecycle tracking using automated processes for clarity, monitoring, and cataloging. AI governance is key to ensuring trust and transparency in AI models. The governance process allows enterprises to direct, manage, and monitor AI activities across business processes. IBM's AI strategy is to provide an end-to-end AI stack for enterprise clients looking for secured AI.

IBM Cloud aims to provide a hybrid cloud AI infrastructure designed for security with:

- Purpose-built datasets for training
- Open models
- Hybrid training and inferencing stacks

Hybrid cloud by design

AI cannot happen without data, but enterprise IT leaders find it difficult to gather the right assets due to complex and siloed IT environments. In particular, GenAI faces challenges linked to distributed or heterogenous environments:

- Multi-model workloads often run on different IT stacks.
- Hybrid multicloud environments require expensive resources.
- Workflows across the enterprise complicate model and data governance.
- Heterogenous environments limit scalability and replicability, and applications remain in the pilot phase for too long.
- Distributed data restricts data quality and access.
- Disjointed and disconnected data can result in lost revenue and security risks while hindering AI productivity.

Solving these challenges requires realizing that hybrid cloud and AI are two sides of the same coin—they cannot be treated as separate concepts. Organizations must be hybrid by design.

> "We're excited to unlock new AI business opportunities for our clients through this collaboration—especially for clients in highly regulated industries. The delivery of Intel Gaudi 3 accelerators on IBM Cloud aims to help clients more cost-effectively test, innovate, and deploy AI inferencing solutions, ultimately scaling enterprise AI with security and performance."
>
> — Satinder Sethi, GM,
> IBM Cloud Infrastructure Services

IBM Cloud empowers enterprises across industries and around the world to leverage the combined power of hybrid cloud and AI by being data-focused, platform-oriented, and AI-infused. Different AI workloads have different requirements, so a single architecture is not appropriate for all workloads. Proper data management and a strategically designed hybrid cloud architecture can help IT leaders make informed choices for their companies so that data is easier to find, and AI is easier to execute.

IBM's intentional approach to hybrid cloud is differentiated by combining cloud expertise with deep hyperscaler and software vendor partnerships, underpinned with Red Hat's open hybrid cloud platform: Red Hat OpenShift.

Flexible deployment options

Customers will have a variety of choices for how they want to deploy their AI workloads on Intel Gaudi 3 accelerators on IBM Cloud:

- **IBM Cloud Virtual Servers for VPC:** This option provides a purpose-built AI server on IBM Cloud Virtual Private Cloud (VPC) for traditional and GenAI workloads, including Red Hat Enterprise Linux (RHEL) AI workloads.

- **Bring-your-own watsonx software license:** Clients requiring added control over their AI stack can deploy IBM watsonx.ai to their Intel Gaudi 3-based virtual server on IBM Cloud VPC in Q2 2025.
- **Deployable architectures:** For developers and operations teams that need to quickly deploy new features and system updates without extensive manual intervention, IBM Cloud offers design modules called deployable architectures (DA). IBM Cloud clients will be able to quickly adopt Intel Gaudi 3 capabilities through several DAs, including a watsonx software DA, an IBM Cloud Virtual Server for VPC DA, as well as DAs for Red Hat OpenShift and Red Hat OpenShift Kubernetes Service. These DAs will be available in 2H 2025.
- **Provision a container worker node:** For clients wanting to leverage managed containerized infrastructure, IBM Cloud plans to deliver Intel Gaudi 3 as a worker node for Red Hat OpenShift AI clusters and Red Hat OpenShift Kubernetes Service in Q2 2025.

Where to Get More Information

Visit these sources for more information:

- Intel® Gaudi® 3 AI accelerators on IBM Cloud
- Intel on IBM Cloud
- Intel® Gaudi® 3 AI accelerators
- watsonx.ai

Solution Ingredients

- Intel® Gaudi® 3 AI accelerators
- Intel® Xeon® processors
- IBM Cloud
- IBM Cloud Virtual Servers for VPC
- IBM watsonx

intel ai

[1] Intel and IBM Collaborate to Provide Better Cost Performance for AI Innovation, IBM Newsroom, August 29, 2024. Intel and IBM Deliver Enterprise AI in the Cloud, Intel Newsroom, August 29, 2024.

[2] Intel Gaudi 3 AI Accelerator whitepaper, October 23, 2024

[3] Creating the future of human resources with watsonx Orchestrate, IBM Cloud case study, October 2022

[4] Setting a new standard for customer support, IBM Cloud case study, October 2023

[5] Generative AI generated 60% of Ansible Playbook Content in IBM CIO Organization Pilot, IBM Cloud case study, October 2023

[6] Test conducted by IBM Research Team in August 2024, with one Intel Gaudi 3 card (Intel Gaudi software driver 1.17.1), IBM granite-8b (FP16) model, optimum-habana inference backend version b6120f8, input size of 1024 tokens, output size of 1024 tokens, and varying number of concurrent requests. Average email is ~100 tokens, estimated based on data from IBM. Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy. Results may vary.

0425/SB/CAT/PDF 364190-001EN