
DELL Technologies
intel
Move Your Business
Upward and Onward
Power your AI initiatives with Dell PowerEdge
XE9680 Servers with Intel® Gaudi® 3 AI Accelerators
Get Started

AI is the new frontier

Artificial intelligence (AI) has the power to unlock innovation and drive outcomes for organizations across industries — and the globe. It's the reason why more businesses are looking to power data-intensive workloads using accelerated computing. But not all AI servers and accelerators are alike. You want server technology that's purpose-built for high-intensity workloads, like machine learning and deep learning (ML/DL) and generative AI (GenAI). And you need high-performance accelerators that complement your specific workloads, workflows and environment.

Dell Technologies and Intel have joined forces to provide an accelerated infrastructure that drives the high-performance results you're looking for but also delivers distinct advantages. Dell PowerEdge XE9680 Servers offer uncompromised AI performance and silicon diversity, while Intel® Gaudi® 3 AI accelerators are efficient, open and trusted to effectively fuel demanding workloads. Together, they help you power AI on your terms.

Read on and learn how Dell Technologies and Intel work jointly to accelerate your AI initiatives.

Learn more

- Web: Dell.com/AI
- DfD Tech Notes: Introducing Dell PowerEdge XE9680 and Intel Gaudi 3 Accelerators
- DfD Tech Notes: PowerEdge XE9680 Rack Integration with Intel Gaudi 3 AI Accelerator

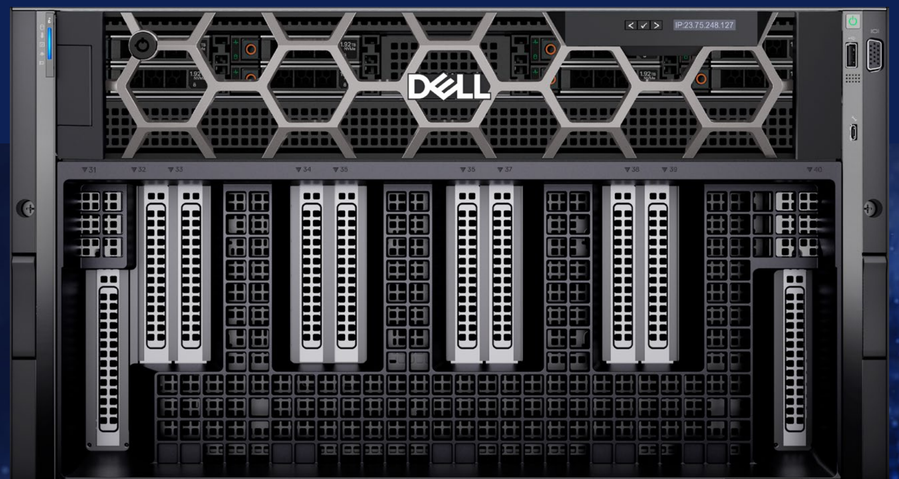
Dell PowerEdge XE9680 Server — for no-compromise accelerated AI

The PowerEdge XE9680 6U Server, the first 8x AI acceleration platform from Dell Technologies, is designed to boost performance for demanding GenAI, ML/DL and high performance computing (HPC) workloads. With up to 64-core 5th Gen Intel Xeon® processors, this platform provides the highest accelerator memory capacity and bandwidth to handle large, complex models and data sets.

Learn more

- Web: PowerEdge XE9680 Rack Server
- Spec sheet: Dell PowerEdge XE9680

PowerEdge XE9680 Server	
Applications and use cases	• GenAI, AI/ML/DL, HPC • Large language models (LLMs), recommendation engines, molecular dynamics and genome sequencing
Processor	2x 5th Generation Intel Xeon Scalable processors
Intel accelerators	Intel Gaudi 3 AI accelerators
Features	• 6U rack height • Air-cooled • 32x DDR5 DIMM slots • Up to 10x 16 PCIe Gen5 slots



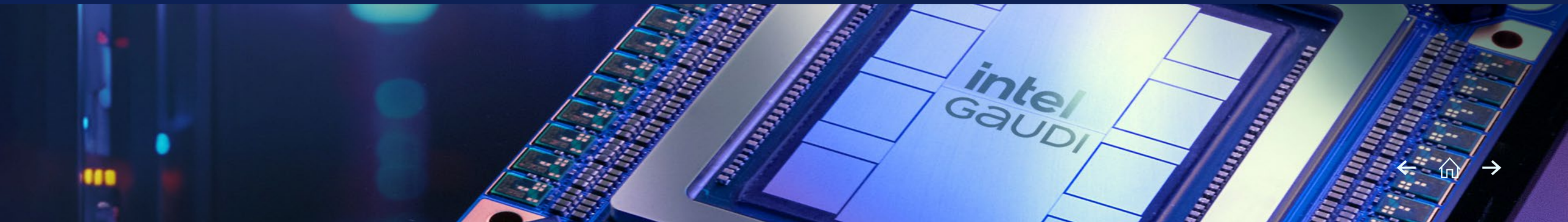
Intel Gaudi 3 AI accelerators — Enterprise-ready performance, scalability and efficiency

Designed for AI and machine learning workloads, Intel Gaudi 3 AI accelerators enable HPC specifically tailored for LLM inferencing and training. Intel Gaudi 3 AI accelerators address customer GenAI needs while reducing TCO and easing deployment via an open software ecosystem and scalable Ethernet-based AI fabrics optimized for PowerEdge.

Learn more

- Technical paper: Intel Gaudi 3 AI Accelerator
- Web: Intel Gaudi 3 AI Accelerators

Intel Gaudi 3 AI accelerators	
Tensor processor cores (5th generation)	64
HBM capacity	128GB
HBM bandwidth	3.7TB/s
MME units	8
Host interface	PCIe Gen5 X16
On-die SRAM capacity	96MB
OAM support	24x 200 GbE RoCE for scale-up and scale-out



Accelerate your AI initiatives with a powerful foundation

As the AI landscape changes and evolves, a resilient and uncompromising infrastructure becomes essential. The PowerEdge XE9680 Server is purpose-built to help you address today's demands and tomorrow's opportunities. It allows you to configure up to eight Intel Gaudi 3 AI accelerators. And, with advanced I/O technologies, it provides a solution that maximizes speed, scalability and performance for the most demanding AI workloads.

No-compromise AI infrastructure

The PowerEdge XE9680 with Intel Gaudi 3 provides advanced capabilities that boost computational performance, accelerate critical applications, and tackle even the most complex workloads with ease.

- **The first Dell 8x AI acceleration platform** enables you to leverage advanced AI capabilities for enhanced computational tasks, leading to improved efficiency and faster project completion.
- **64-core 5th Gen Intel Xeon processors** deliver powerful performance, enabling faster processing and improved efficiency for critical applications.
- **Large accelerator memory and bandwidth** put you in a position to manage large and complex models and data sets, essential for big data and AI workloads.

Tailored to your needs

Configure your infrastructure in a manner that fits your environment and effectively powers your workloads. Silicon diversity in the PowerEdge XE9680 gives you freedom to configure your foundation with Intel Gaudi 3 AI accelerators.

- **8 Intel Gaudi 3 OAM accelerators** enable you to scale performance to meet specific workload demands. Drive optimal resource allocation to support growth and capitalize on more opportunities.
- **Ethernet connectivity with embedded RoCE ports.** Enhance data transfer speeds in situations that require low-latency connections for real-time applications.
- **1.5TB shared coherent memory** supports enhanced GenAI training performance, enabling you to innovate and deploy AI solutions faster.

Accelerated I/O throughput

Don't let anything get in the way of real-time insight generation — especially bottlenecks. Drive high performance and reduced latency for data-intensive workloads with an AI infrastructure that prioritizes and accelerates I/O throughput.

- **DDR5, PCIe Gen 5.0 and NVMe® SSDs** push data flow and computing possibilities. Maximize your computing efficiency for faster product development cycles and market responsiveness.
- **8 front-facing PCIe Gen 5 slots** offer optimal expansion for real-time AI operations — and the flexibility to adapt to changing technology needs.
- **16-drive capacity** provides enhanced storage for high-performance tasks, ensuring that your business can handle demanding applications without bottlenecks.

Learn more

- Infographic: Go boldly onward with Dell PowerEdge XE9680 and Intel Gaudi 3

Comprehensive solutions with Dell AI Factory

These end-to-end AI solutions combine the power of Intel Gaudi 3 AI accelerators with Dell servers, storage, networking, services and software to provide jointly engineered, tested and validated solutions to ensure a seamless deployment experience, all on your terms. Proven with Dell Validated Designs, these solutions can help you get started on your journey with AI use cases that help you accelerate your business outcomes.

Customer benefits

1. High performance

Intel Gaudi 3 AI accelerators showcase exceptional MLPerf performance in vLLM model serving, delivering high throughput and scalability. At 256 concurrent users, the system outputs 2,302 tokens/sec, far surpassing the 10 tokens/sec industry benchmark for real-time AI. As concurrency increases, performance scales impressively, reaching 5,142 tokens/sec at 1,024 users and 5,236 tokens/sec at 4,096 users,[1] making Gaudi 3 ideal for latency-sensitive, high-demand AI applications like chatbots.

2. Scalability

Designed for growth, these solutions adapt to expanding data, user needs and advanced AI use cases without requiring major infrastructure changes, keeping businesses agile and future-ready.

3. Flexibility

An open Ethernet environment, open-source software stack, and Dell Validated Designs provide customization, ease of deployment and enhanced productivity. With cutting-edge networking options like Dell Enterprise SONiC and PowerSwitch Z9864F-ON, businesses can tailor solutions to meet their AI goals effectively.

[1] Dell Blog, AI Agents Revolutionizing Customer Service: Agentic Multi-Modal RAG Solution Powered by Dell PowerEdge™ XE9680, January 2025.

End-to-end solution components.

Simplify and **accelerate** your AI journey with the right technologies, services and partners.

Create better outcomes with AI solutions **tailored** to your unique business.

Protect and sustain your success on a **trusted** foundation that keeps you in control.

Content creation

Digital assistants

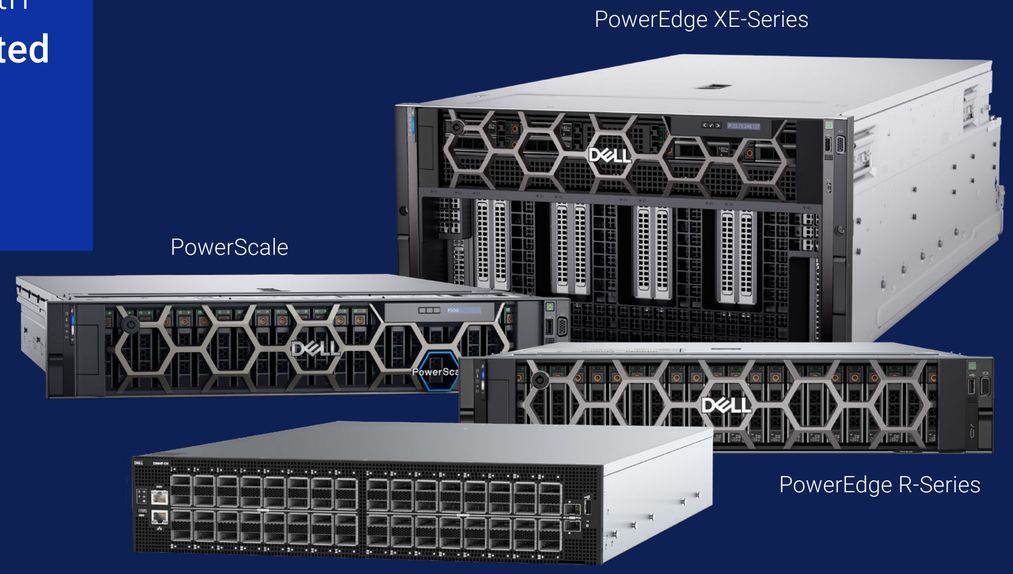
Code generation

Design and data creation

AI use cases

Proven with **Dell Validated Designs**

Partner with **Dell Services**



intel. XEON

intel. GAUDI

Networking · Compute · Storage

Fully validated open-source solution

Our joint solutions take full advantage of an open-source software stack to boost performance and accessibility across environments. This stack includes Intel Gaudi Software Suite and integrates tools like PyTorch® and Hugging Face®, simplifying the fine-tuning, training and deployment of AI models. Each component is optimized to enhance efficiency and flexibility for developers.

A key feature is Dell Omnia, an open-source tool for deploying and managing high-performance clusters tailored for HPC, AI and data analytics. Dell Omnia streamlines Kubernetes® installation and resource allocation, enabling IT teams to efficiently deploy and manage diverse workloads.

OMNIA

Tools and models	Hugging Face, Meta® Llama® 3
ML frameworks	PyTorch
ML tools	Jupyter, Kubeflow
ML model serving	vLLM, TGI, TEI
Observability	Grafana®, Prometheus, Metrics: Intel Gaudi 3, Kubernetes, iDRAC, Dell PowerScale OneFS
Optimized libraries Development environment Programming models Drivers and runtime	Intel Gaudi Software Suite, Optimum
Workload manager	Kubernetes
Operating system	Ubuntu

Dell PowerEdge (iDRAC, BIOS, firmware) configuration

Experience the advantage

The XE9680 Server with Intel Gaudi 3 AI accelerators drives high performance for every type of enterprise AI workload. Together they also offer distinct advantages — from simple, efficient networking and platform scalability to faster ROI and the backing you can only get from a trusted partnership.

#1 Efficient networking

The PowerEdge XE9680 Server, combined with Intel Gaudi 3 AI accelerators, drives networking efficiency by reducing complexity and lowering total cost of ownership.

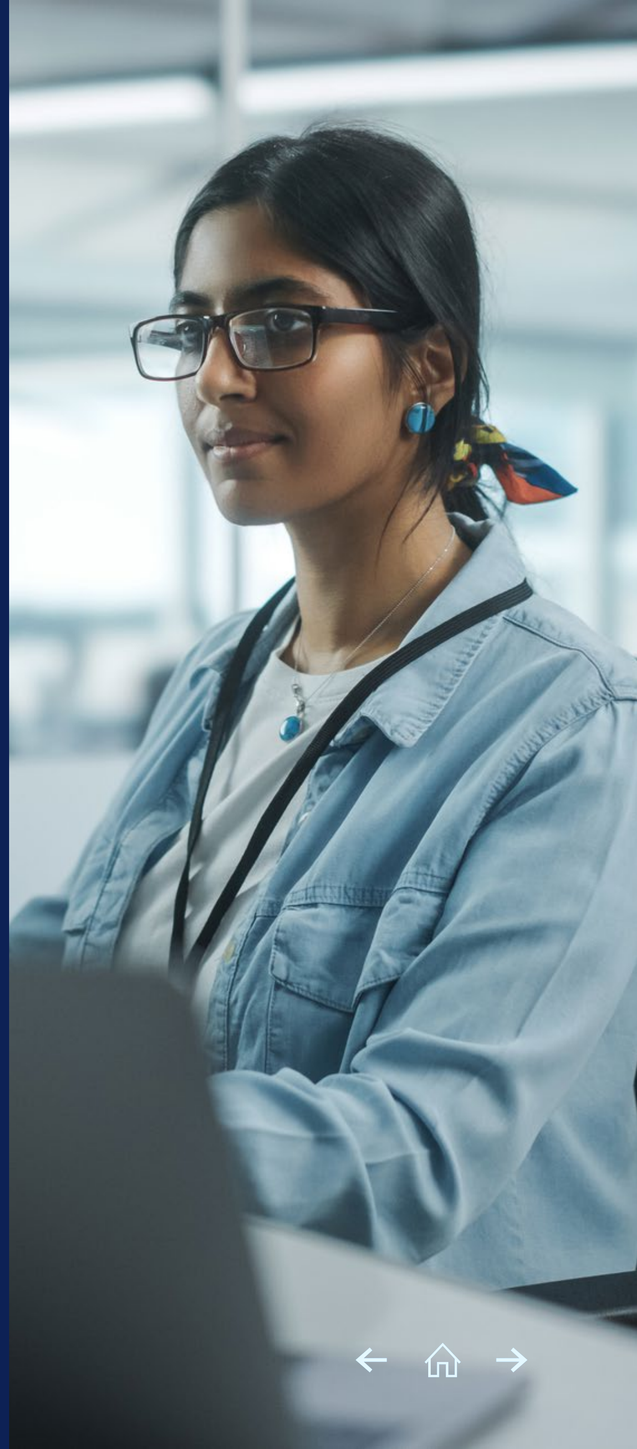
- The XE9680 includes 6 OSFP 800GbE ports.
- Each Intel Gaudi 3 AI accelerator is equipped with 24 integrated 200 GbE ports.

With integrated networking that connects directly to an external accelerator fabric, you eliminate the need for external NICs.

#2 Scalability

Easily manage growing data volumes and complex tasks with a highly scalable infrastructure. This adaptable system ensures high performance and operational efficiency, enabling seamless growth well into the future.

- Intel Gaudi 3 is designed to support demanding GenAI workloads, scaling from one to thousands of nodes.
- The PowerEdge XE9680 offers up to 32 DDR5 memory DIMM slots and eight U.2 drives.



#3 Open software ecosystem

Simplify development with an open ecosystem and easy migration.

- Integrated open-source PyTorch framework with optimized model library on Hugging Face
- Migrate models on open software with as few as three lines of code.

#4 Fast ROI

Accelerate the return on your investment with a scalable, efficient and open foundation. Easily and efficiently accommodate rising data volumes and future growth with a resilient AI infrastructure that's built to scale.

- Lower deployment TCO with standard Ethernet switches and fewer NICs.
- Boost efficiency with air-cooled PowerEdge servers and Dell Integrated Rack Scalable Systems (IRSS).
- Extend your ROI over time with long-life Intel Gaudi 3 AI accelerators.

#5 Partnership

Dell Technologies and Intel are committed to driving your business success with powerful, reliable technology solutions. For years, we've partnered with organizations like yours to deliver essential outcomes through trusted Dell and Intel systems. Looking ahead, we'll continue to support your technology and business needs with future-ready solutions that help you seize the next big trends.

Read our blog articles:

- Agentic RAG solution powered by Dell PowerEdge and Intel Gaudi 3
- Enterprise AI Deployment with Dell PowerEdge and Intel Gaudi 3

Accelerate your journey to insight and innovation

Whether you're running AI, ML/DL or GenAI workloads to drive new insights or enhance operations, start with a solid foundation. PowerEdge XE9680 Servers and Intel Gaudi 3 AI accelerators provide the high performance results you demand with efficiency, scalability — and the flexibility of an open ecosystem. Trust Dell Technologies, your innovation catalyst, and Intel to elevate your business, taking it upward and onward into the realm of new possibilities.

Learn more at Dell.com/AI