
DELL Technologies
intel GAUDI
Move Your Business
Upward and Onward
Power your AI initiatives with Dell PowerEdge XE9680 and XE7740 servers with Intel® Gaudi® 3 AI Accelerators
Get Started

AI is the new frontier

AI is transforming industries worldwide — but gaining a real edge requires the right foundation. Data-intensive workloads like ML/DL and GenAI demand servers built for performance and AI accelerators optimized for your unique workflows. That's how you're going to drive faster innovation, greater efficiency and measurable business impact.

Dell Technologies and Intel have joined forces to deliver the accelerated infrastructure you need. Dell PowerEdge XE9680 and XE7740 Servers provide the performance AI workloads demand, while Intel® Gaudi® 3 AI accelerators offer an efficient, open and trusted way to fuel even the most intensive tasks. Together, they give you the freedom to power AI on your terms.

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

Learn more

- Web: Dell.com/AI
- Web: Intel Gaudi 3 Accelerator
- 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

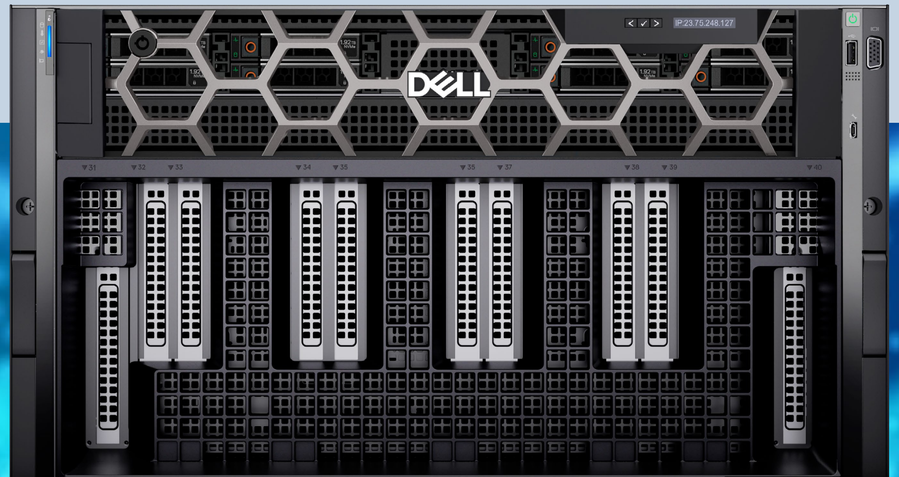
Dell PowerEdge XE9680 Server — for no-compromise accelerated AI

The PowerEdge XE9680 server, the first 6U server with 8x AI acceleration from Dell Technologies, is designed to boost performance for demanding AI and high performance computing (HPC) workloads. With two 5th Gen Intel Xeon® processors, with up to 64 cores per processor, 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	• AI inferencing and fine-tuning • 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 • 10x PCIe Gen 5 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 AI specifically tailored for LLM inferencing. 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
High Bandwidth Memory (HBM) capacity	128GB
HBM bandwidth	3.7TB/s
Matrix Multiplication Engine (MME) units	8
Host interface	PCIe Gen 5 x 16
On-die SRAM capacity	96MB
OCP Accelerator Module support	24x 200 GbE RoCE for scale-up and scale-out

POWEREDGE XE9680

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 6U server with 8x AI acceleration** enables you to leverage advanced AI capabilities for enhanced computational tasks, leading to improved efficiency and faster project completion.
- **64-core Intel 5th Gen 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.

- **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 inferencing and fine-tuning 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 memory and NVMe® SSDs** push data flow and computing possibilities. Maximize your computing efficiency for faster product development cycles and market responsiveness.
- **10 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

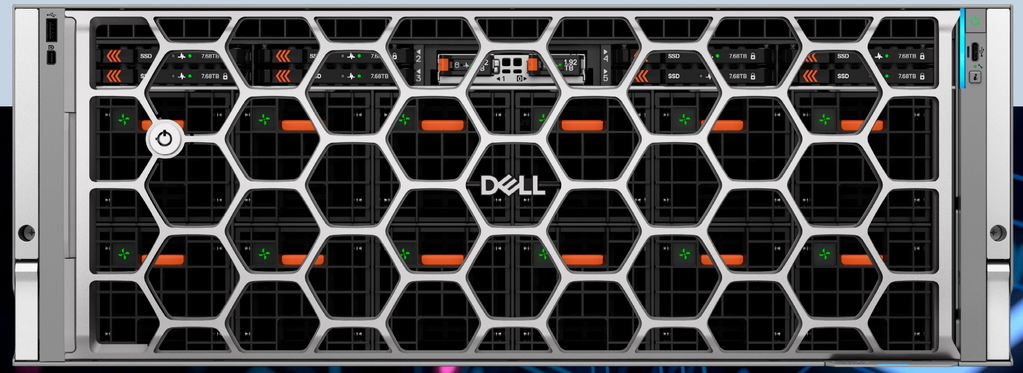
Dell PowerEdge XE7740 Server — for flexible AI acceleration

The Dell PowerEdge XE7740 is purpose-built for large-scale AI inferencing, delivering powerful acceleration, GPU flexibility, advanced networking and robust security in a 4U scalable design. Ideal for industries from finance to healthcare, it simplifies on-prem deployment, drives efficiency and sets a new benchmark for enterprise AI.

Dell PowerEdge XE7740 Server

- Web: PowerEdge XE7740 Rack Server
- Spec sheet: Dell PowerEdge XE7740
- Infographic: Aim high with the right server for AI

Dell PowerEdge XE7740	
Applications and use cases	• AI inferencing • AI model fine-tuning • AI-powered HPC applications
Processor	2x Intel Xeon 6 P-core processors with up to 86 cores per processor
Intel accelerators	Intel Gaudi 3 AI Accelerator PCIe cards
Features	• 4U rack server • Air cooled • 32 DDR5 DIMM slots • 4- or 8-card PCIe configurations

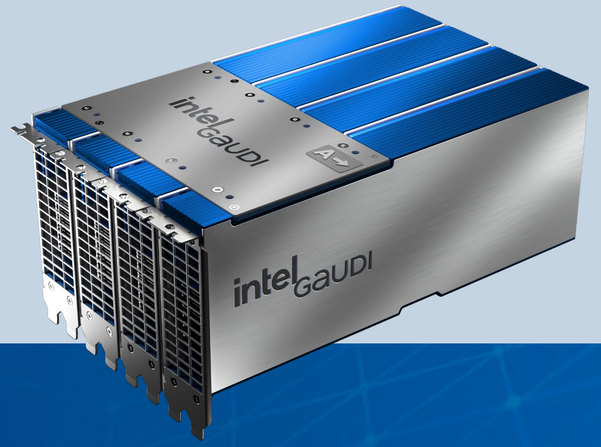


Intel Gaudi 3 AI Accelerator PCIe card — scalable, cost-efficient AI performance in a standard rack form factor

The Intel Gaudi 3 AI Accelerator PCIe card is designed to provide scalable, cost-efficient AI performance for inferencing and fine-tuning of today's most widely used models. It delivers a familiar, easy-to-deploy server hardware form factor that helps enterprises accelerate innovation with powerful yet economical AI acceleration.

Learn more

- Product brief: Intel Gaudi 3 AI Accelerator HL-338 PCIe Card



Intel Gaudi 3 AI Accelerator PCIe card	
Architecture	5th generation tensor processor core
High Bandwidth Memory (HBM) capacity	128GB
HBM bandwidth	3.7TB/s
Thermal Design Power (TDP)	600W (air cooling)
On-die SRAM capacity	96MB
Scale-out support	Via Host-NIC
Gaudi 3 PCIe accelerator interconnectivity	Up to 2 groups of 4-way bridge cards for direct High Bandwidth Memory sharing — bypassing PCIe slots

POWEREDGE XE7740

Power AI innovation with confidence

The Dell PowerEdge XE7740 with Intel Gaudi 3 AI Accelerator PCIe cards delivers enterprise-grade performance and flexibility, making it easy to integrate advanced AI into existing infrastructure. Together, we accelerate innovation, optimize costs, and overcome power and scalability challenges — while preparing your business for the future.

Flexible, high-performance AI

Experience the freedom to design your right-sized AI infrastructure.

- **Configurable, high-density GPU options.** Choose 4 or 8 double-wide PCIe cards, optimized for virtually any data center rack power profile.
- **Air-cooled 4U rack integration.** Support for up to 8x 600w PCIe AI accelerators.
- **One-to-one GPU-to-NIC ratio.** Avoid vendor lock-in and optimize throughput with the flexibility to choose your network interface.
- **Massive memory capacity.** Each accelerator card comes with 128GB HBM2e memory and 96MB on-die SRAM — ideal for running low-latency inferencing and LLM workloads.
- **Exceptional value.** Offers up to 3x more performance per dollar than competing GPUs.[2]

[2] See backup for workloads and configurations. Your costs and results may vary.

Seamless integration

Simplify deployment and maintenance with this solution's air-cooled 4U chassis and standard PCIe form factor.

- **Easy to integrate — no retrofits or modifications needed.** The air-cooled XE7740 fits standard racks, enabling fast, hassle-free deployment for on-premises AI workloads.
- **Maximize GPU density.** Pack more PCIe GPUs per rack without compromising airflow.
- **Built for growth.** Eight front-serviceable PCIe slots and an integrated OCP 3.0 Ethernet module allow you to scale bandwidth as AI demand rises.
- **Native software support.** Intel Gaudi 3 accelerators integrate directly with PyTorch, vLLM and Hugging Face, making it easy to fine-tune and serve models in production.

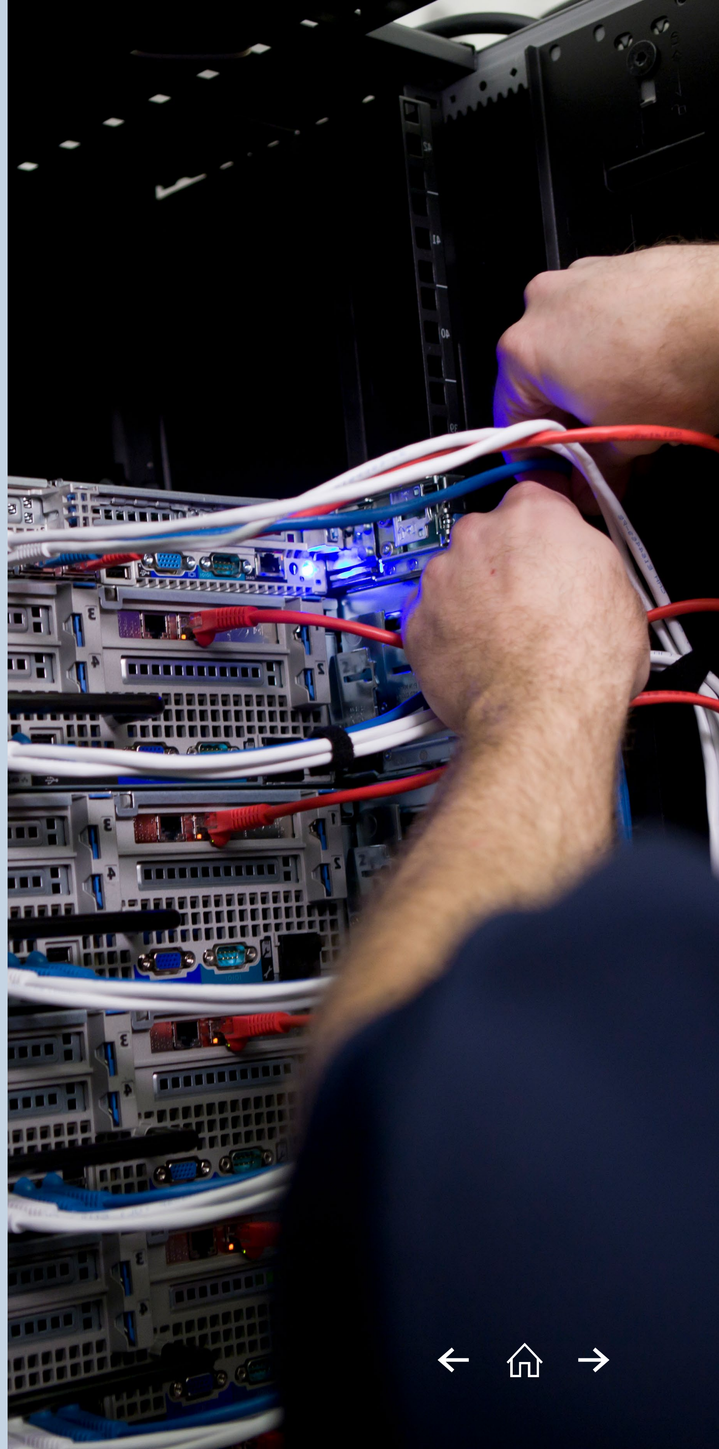
Scalable, open and secure

Work with a solution that's built for today and tomorrow.

- **Open ecosystem.** Dell PowerEdge and Gaudi 3 support a truly open architecture, eliminating vendor lock-in.
- **Zero Trust security.** Protected by silicon root of trust, cryptographic verification and firmware safeguards.
- **Simplified management.** Dell OpenManage automates deployment, monitoring and updates across clusters.

Learn more

- Blog: PowerEdge XE7740 with Gaudi 3 breaks barriers to enterprise AI accessibility



Experience the advantage

The PowerEdge XE9680 and XE7740 servers 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.

The XE7740 offers highly flexible networking, supporting up to a 1:1 accelerator-to-NIC ratio with eight full-height PCIe slots and an integrated OCP networking module.

#2 Scalability

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

- Intel Gaudi 3 AI accelerators are designed to support demanding AI workloads, scaling from one to thousands of nodes.
- Intel Gaudi 3 PCIe cards provide a modular AI growth path without overcommitment, making them a great choice for early pilots or phased AI deployments.

#3 Open software ecosystem

Simplify development with an open ecosystem and easy migration.

- Intel Gaudi 3 AI accelerators support an open ecosystem, allowing AI developers to innovate freely without vendor lock-in, ensuring flexibility for future advancements.
- 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.

- The XE7740 offers a compelling price-to-performance ratio, ensuring you get more AI capability for less.
- Lower deployment TCO with standard Ethernet switches and fewer NICs.
- Boost efficiency with air-cooled PowerEdge servers.

#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, innovation and growth.

Whether you're running AI, ML/DL or GenAI workloads to drive new insights or enhance operations, start with a solid foundation. PowerEdge XE9680 and XE7740 servers with 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 about PowerEdge XE9680 and XE7740 servers.