intel.

DELLTechnologies

intel
partner   Titanium

Empowering GenAI Choice: Dell AI Factory Expands with Intel

By Chad Dunn, Vice President of Product Management for Artificial Intelligence and Data Management, Dell Technologies

Meet Dell Generative AI Solutions with Intel, a new expansion to the Dell AI Factory, featuring Intel® Gaudi® 3 AI Accelerators.

AI Solutions Generative AI Partners.

As generative AI (GenAI) adoption continues to accelerate in virtually every industry, organizations face critical challenges such as ensuring data privacy and security, identifying the right infrastructure, tools, models and platforms for their use cases, right sizing their AI investments, seamlessly integrating new AI technologies with existing systems and overcoming the shortage of highly skilled personnel. Dell Technologies is responding to these challenges with the introduction of Dell Generative AI Solutions with Intel, an exciting expansion of the Dell AI Factory.

Set for general availability by year end, these solutions are designed to accelerate AI innovation by combining the power of Intel® Gaudi 3® AI accelerators with Dell's high-performance servers, storage, networking, professional services and a full stack of open-source software. These jointly engineered solutions offer high performance at an optimized TCO, massive scalability and the flexibility to choose the right silicon, ethernet, software and implementation model for your GenAI workflows.
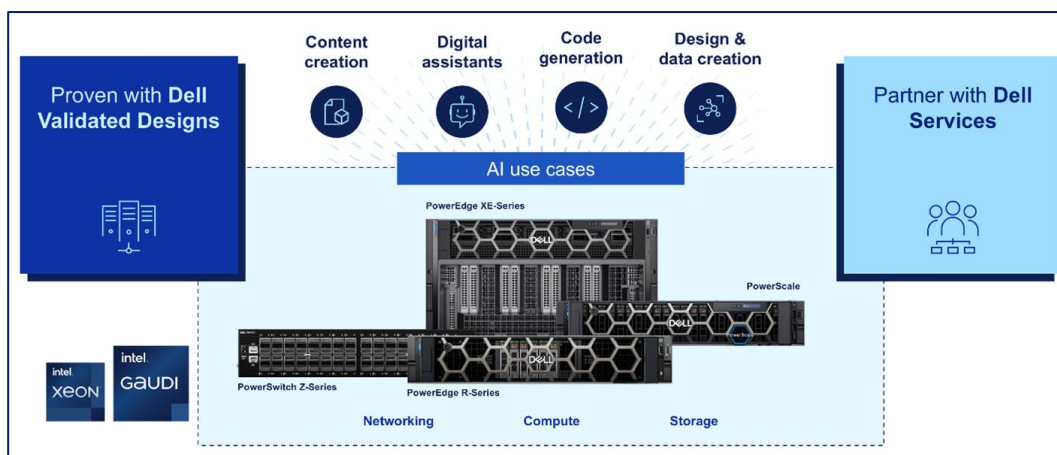

Fig. 1 Dell Generative AI Solutions with Intel components.

Key Features and Benefits

1. **High Performance**

Dell Generative AI Solutions with Intel are engineered to provide excellent performance while reducing total cost of ownership (TCO). These solutions include Dell PowerEdge XE9680s equipped with an Intel® Gaudi® 3 baseboard (HLB-325) featuring 8 Intel® Gaudi® 3 AI Accelerator OAM mezzanine cards, and dual 5th Generation Intel® Xeon® Scalable processors.

Intel Gaudi® 3's built-in Ethernet links offer 200Gbps each. 8 x Gaudi® 3 modules can reach a theoretical peak bandwidth of 4,200GB/s eliminating the need for external NICs on the back-end while providing a higher aggregate bandwidth than proprietary alternatives. Paired with Intel E810 NICs on the front-end, the system ensures 10/25GbE speeds, advanced packet filtering and RoCEv2 support, delivering scalable and high-performance networking for external latency-sensitive data traffic. [New Gaudi® 3 AI accelerators](#) offer up to 20% more throughput and 2x price/performance for inference of Llama 2 70B vs. the leading competitor, empowering businesses to enhance their AI workloads without compromising efficiency. See how [Intel® Gaudi® 3 and Dell PowerEdge XE9680 work together to support demanding AI initiatives.](#)

These solutions also leverage [Dell PowerScale All-Flash Storage](#) and [Dell PowerSwitch networking](#), all engineered for demanding AI workloads.

2. **Scalability**

Future growth and changes in GenAI initiative is inevitable. Our solutions provide both massive scalability and flexibility, allowing organizations to support immediate objectives today, while strategically positioning themselves for future growth. This means that whether you're adding new users, processing more data or implementing advanced use cases, your infrastructure can adapt without requiring a complete overhaul, ensuring you remain agile and competitive in a rapidly changing landscape.

3. **Flexibility**

An open standard Ethernet networking environment, open-source software stack and pre-validated implementation options empower you to deploy the solution that best supports your GenAI journey and enhances developer productivity. Take advantage of a standards-based AI fabric along with [Dell Enterprise SONiC](#) and Dell PowerSwitch Z9864F-ON, featuring 64 ports of 800G connectivity. Dell GenAI Solutions with Intel are proven by Dell Validated Design, making it easy to right-size your investment, without the guesswork.

Fully Validated Open-source Ecosystem

Dell Generative AI Solutions with Intel are designed using an open-source software stack to optimize performance and accessibility across diverse environments. This stack incorporates Intel® Gaudi® software and ensures seamless interoperability with a wide range of industry-leading open-source tools, libraries and models. Included in this comprehensive ecosystem are platforms, runtimes, frameworks, extensions and libraries such as PyTorch, Hugging Face, Meta, vLLM, TGI, TEI and more, making it easier to find, fine-tune, train, deploy, monitor and serve GenAI models at scale. Each component has been carefully selected and validated to maximize both efficiency and flexibility in AI model development and deployment, empowering developers to innovate rapidly while maintaining high performance.

Central to the solutions is Dell Omnia, an open-source software orchestrator designed for the deployment and management of high-performance clusters tailored for HPC, AI and data analytics workloads. Omnia facilitates the installation of Kubernetes for efficient job management, while also enabling the integration of various packages and services to support diverse workloads within a single converged solution. Developers are actively enhancing Omnia to accelerate the deployment of new infrastructure into resource pools that can be easily allocated and reallocated for different workloads. By streamlining this process, Omnia empowers IT teams to deliver the right tools for each job on the appropriate infrastructure when needed.
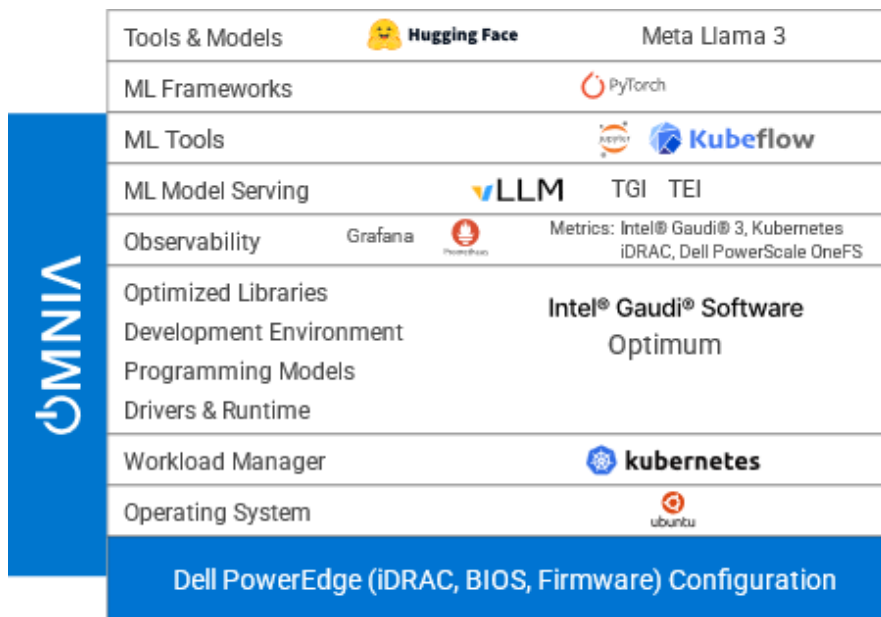

Fig. 2 Dell Generative AI

Get Started with Dell Professional Services for Generative AI

Dell Professional Services provides an extensive portfolio of professional services as part of the Dell AI Factory to support every stage of your generative AI (GenAI) journey. We work alongside you to build a strategy and roadmap, validate and securely managed data to power AI, implement, train, validate and support GenAI models and simplify operations close skills gaps for maximized ROI of AI solutions. Learn more about how Dell Professional Services can help you get started on your Generative AI journey.

Ready to explore the potential of AI in your enterprise? Sign up for a fee-waived Accelerator Workshop for Generative AI. This workshop helps gain consensus among business and technical stakeholders on your solution and prioritized use cases.

Learn More About Dell AI Solutions

Contact a Dell Technologies Expert

Stay ahead in the AI race with Dell Generative AI Solutions with Intel. Join the conversation and transform your business operations with cutting-edge AI technologies.

Company Name	Solution Name	Booth Info
Dell Technologies™	Dell AI Factory™	Booth #5721