Dell Technologies Expands Dell AI Factory with New PowerEdge Servers to Accelerate Enterprise AI Adoption

October 10, 2024

Additions to the world's broadest GenAI solutions portfolio empower organizations with flexible AI solutions

SAN FRANCISCO, Oct. 10, 2024 /PRNewswire/ --

News Summary

- Additions to the Dell AI Factory span infrastructure, solutions and services to simplify and speed enterprise AI adoption
- Five new Dell PowerEdge servers with AMD 5th Generation EPYC processors deliver industry-leading AI performance, efficiency and flexibility at scale
- Dell Generative AI Solutions for AMD simplify AI deployment with significant time-to-value improvements
- The Dell Enterprise Hub now supports Dell PowerEdge XE9680 with AMD Instinct™ MI300X accelerators to securely deploy the latest Llama and Mixtral models
- Dell Implementation Services for Generative AI Platform provides easy to manage, secure on-premises AI platform solutions to speed AI innovation

Full story

AMD Advancing AI – Dell Technologies (NYSE: DELL) expands the world's broadest generative AI (GenAI) solutions portfolio with Dell AI Factory additions tailored for AMD environments. These solutions offer enterprises enhanced AI capabilities, including greater scalability and flexibility, to stay competitive in the evolving technology landscape.

"By integrating AMD technology into the latest Dell servers, AI solutions and services through the Dell AI Factory, we're providing the performance and efficiencies enterprises need today and in the future," said Arthur Lewis, president, Infrastructure Solutions Group, Dell Technologies. "Together with AMD, we are setting new standards in AI performance, giving enterprises powerful and cost-effective solutions essential for modern data-driven environments."

"Our latest AMD EPYC processors and Instinct accelerators, provide Dell customers with advanced end-to-end solutions, at scale, that will speed AI deployments and power business critical IT infrastructure," said Forrest Norrod, executive vice president and general manager, Data Center Solutions Group, AMD. "Our longstanding collaboration with Dell continues to deliver cutting edge solutions that help enhance operational efficiency and reduce time to value, fostering innovation across highly competitive industries."

New Dell PowerEdge servers drive demanding AI workloads at scale

The additions to the Dell PowerEdge portfolio drive a wide range of AI use cases and traditional workloads and simplify server management and security. The platforms provide customizable and efficient solutions that simplify management and support high-performance workloads for modern enterprises:

- Designed for enterprise AI workloads, the **Dell PowerEdge XE7745** supports up to eight double-width or 16 single-width PCIe GPUs with AMD 5th Generation EPYC processors in a 4U air-cooled chassis. Purpose-built for AI inferencing, model fine-tuning and high performance computing, the internal GPU slots are paired with eight additional Gen 5.0 PCIe slots for network connectivity, creating dense, flexible configurations with 2x more DW PCIe GPU capacity.[1]
- The **PowerEdge R6725 and R7725** servers are optimized for scalability with high performing AMD 5th Generation EPYC processors. The new DC-MHS chassis design enables enhanced air cooling and dual 500W CPUs, conquering tough thermal challenges for power and efficiency.[2] These platforms maintain rigorous data analytics and AI workloads, with configurations optimized for scalability, and offer record-breaking performance for workloads like virtualization, databases and AI.[3] The R7725 offers up to 66% increased performance and up to 33% increased efficiency at the top of the stack.[4]

All three platforms can support up to 50% more cores, with up to 37% increased performance per core resulting in greater performance, efficiency and improved TCO.[5] These gains consolidate up to seven 5-year-old servers into one server today, resulting in up to 65% lower CPU power consumption.[6]

- The **PowerEdge R6715 and R7715** servers with AMD 5th Gen EPYC processors offer increased performance, efficiency and up to 37% increased drive capacity resulting in greater storage density.[7] Available in various configuration options, the

single-socket servers support double the memory with support for 24 DIMMs (2DPC), and meet diverse workload requirements and maximize performance in compact 1U and 2U chassis.[8] The R6715 sees world record performance for AI and virtualization tasks.[9]

For customers deploying AI at scale, Dell Technologies will also continue to support all the latest AMD Instinct accelerators in Dell PowerEdge XE servers.

IT teams can remotely monitor, manage and update Dell PowerEdge servers with the updated Integrated Dell Remote Access Controller (iDRAC). With a faster processor, increased memory and dedicated security co-processor, iDRAC simplifies server management and security, allowing IT teams to respond with greater reliability and efficiency.

"The systems provided by Dell Technologies and AMD for OSF Healthcare allow us to deliver better services to our clinicians and patients, reduce our overall cost and help communities in need. When you have patient lives dependent on our platforms, it is critical that our systems remain stable and operational 24/7, 365 days a year," said Joe Morrow, director, Technology Services, OSF Healthcare. "Because of these systems, we have significantly reduced Epic downtimes, empowering OSF Healthcare to provide superior healthcare services while ensuring security and scalability in our operations."

Dell AI Factory enhancements simplify and speed GenAI deployments

The Dell AI Factory is designed to support diverse AI requirements with the world's broadest AI solutions portfolio from desktop to data center to cloud, allowing organizations to right-size their AI investments:

- **Dell Generative AI Solutions with AMD** using the Dell PowerEdge XE9680 server with AMD Instinct™ MI300X accelerators to deliver AI inferencing, retrieval augmented generation (RAG) and customization. The integration simplifies AI deployment, enhances security, allows for scalable and modular architectures and reduces time to value by up to 86%, helping organizations optimize their AI investments.[10]
- The **Dell Enterprise Hub** on Hugging Face now supports the Dell PowerEdge XE9680 with AMD Instinct MI300X accelerators, providing custom containers and scripts for easy, secure deployment of models such as Llama and Mixtral. These containerized models are uniquely optimized to boost inferencing performance based on the model and server and leverage the Hugging Face Text Generation Inference (TGI) back-end.
- Dell Professional Services for Generative AI are expanding to support AMD environments. New **Dell Implementation Services for GenAI Platform** with AMD provide a tailored operational platform, including Kubernetes configuration, deployment of advanced AI frameworks and customer knowledge transfer on best practices to maximize ROI. From strategy development to data preparation, platform and model implementation and solution scaling, Dell professionals help customers achieve business outcomes with AI throughout the lifecycle.

"Dell Technologies and AMD continue to drive AI innovation, offering comprehensive solutions and services that help businesses modernize data centers, improve scalability, and leverage AI for better business outcomes," said Kuba Stolarski, Research VP, IDC. "The latest updates to the Dell AI Factory are the next step in Dell's journey of delivering AI solutions that boost efficiency and value."

Availability

- The Dell PowerEdge XE7745 server will be available globally starting January 2025.
- The Dell PowerEdge R6715, R7715, R6725 and R7725 servers will be available globally starting November 2024.
- Dell Generative AI Solutions with AMD will be available globally starting in Q4 2024.
- Dell Enterprise Hub on Hugging Face updates are available globally now.
- The Dell Implementation Services for Generative AI platform is available in select countries now.

Additional Resources

- Connect with Dell on X and LinkedIn
- Find out more about Dell Generative AI Solutions with AMD
- Visit the Dell Enterprise Hub on Hugging Face
- Learn more about AI at Dell

About Dell Technologies
Dell Technologies (NYSE: DELL) helps organizations and individuals build their digital future and transform how they work, live and play. The company provides customers with the industry's broadest and most innovative technology and services portfolio for the AI era.

1 Based on Dell analysis of the Dell PowerEdge XE7745 specification sheet showcasing 8x PCIe Gen5 slots in addition to the GPU slots. Additionally, comparing the XE7745 with the prior generation Dell PowerEdge R760XA, the XE7745 has support for up to 8 DW or 16SW GPUs while the R760XA has support for up to 4 DW or 12 SW GPUs. Data accurate as of 10/2/2024.

2 Based on Dell analysis comparing the top CPU supported in a Dell PowerEdge R7725 of the AMD EPYC 5th gen CPU with 192 zen 5c cores to the top CPU supported in the Dell PowerEdge R7625 of the AMD EPYC 4th gen CPUs with 128 zen 4c cores. Additionally, based on specification data for the Dell PowerEdge R7725 showing support for 2x 500W CPUs with air cooling. Data accurate as of 10/2/2024. Actual performance might vary.

3 Based on Dell PowerEdge servers achieving world record scores for VMMark4 (2-S matched pair score of 5.17@5.8 tiles with the Dell PowerEdge R7725), SAP-SD (2-S score of 201,000 users with the Dell PowerEdge R6725), and TPCx-AI (Scores of 720.386 @SF3, 864.593@SF10 and 1115.609@SF30 with the Dell PowerEdge R6715 and R6725) as of 10/2/2024. Actual performance might vary.

4 Based on Dell analysis of the SPECIntRate scores of the AMD EPYC 5th gen 9965 CPU of 2980 in the R7725 with that of the AMD EPYC 4th gen 9754 CPU of 1790 in an R7625. Data accurate as of 10/2/2024. Actual performance will vary.

Based on Dell analysis of the SPECIntRate scores of the AMD EPYC 5th gen 9965 CPU of 2980 which has a cTDP of 500W in the R7725 with that of the AMD EPYC 4th gen 9754 CPU of 1790 which has a cTDP of 400W in an R7625. Efficiency for each config is calculated by dividing the SPECInt Score with the default TDP. Data accurate as of 10/2/2024. Actual performance will vary.

5 Based on Dell analysis comparing the top SKU supported in a Dell PowerEdge R7725 of the AMD EPYC 5th gen CPU with 192 zen 5c cores to that supported in the Dell PowerEdge R7625 of the AMD EPYC 4th gen CPUs with 128 zen 4c cores to the top SKU. Data accurate as of 10/2/2024. Actual performance might vary.

Based on Dell analysis of the SPECFPRate scores of the AMD EPYC 5th gen 9755 CPU of 2270 in the R7725with that of the AMD EPYC 4th gen 9754 CPU of 1420 in an R7625. Data accurate as of 10/2/2024. Actual performance will vary.

6 Based on Dell analysis comparing the SPECint and SPECFP scores of the AMD EPYC 5th Gen 9965 in a Dell R7725 (2980 and 2350) with the same scores for an Intel Xeon 8280 in a Dell PowerEdge R740XD (375 and 296). The ratio of the scores shows that 7 of the R740xd servers would give a total score similar to that for the single R7725 as configured above. The CPUs in a single R7725 would have a total TDP of 1000W (2x500W). The CPUs in 7x R740XDs would have a total TDP of 2870W (2*205*7) where each Intel Xeon 8280 has a TDP of 205W. This represents a CPU power reduction of 65%. Data accurate as of 10/2/2024. Actual performance will vary.

7 Based on Dell analysis of specifications comparing the Dell PowerEdge R67x5 server with up to 22 E3.s drive slots with the Dell PowerEdge R66x5 servers with up to 16 E3.s drive slots. Data was collected as of 10/2/2024.

8 Based on Dell analysis of the specifications of the Dell Rx715 which have 24 DIMM slots compared to the Dell Rx615 servers which have 12 DIMM slots each. Data accurate as of 10/2/2024.

9 Based on Dell PowerEdge servers achieving world record scores for SPECVirt (1-S 4-node SAN score of 3.38 with the Dell PowerEdge R6715 is a world record score for SPECVirt at the 32 core per system level), and TPCx-AI (Scores of 720.386 @SF3, 864.593@SF10 with the Dell PowerEdge R6715 are world records for performance) as of 10/2/2024. Actual performance might vary.

10 Estimate based on Dell analysis in May 2024 comparing time to setup a 2-node Kubernetes cluster for a general-purpose LLM using automated scripts vs deploying a common design manually. Setup time includes base installation only. Actual setup time will vary depending on solution configuration.

View original content to download multimedia:https://www.prnewswire.com/news-releases/dell-technologies-expands-dell-ai-factory-with-new-poweredge-servers-to-accelerate-enterprise-ai-adoption-302273138.html

SOURCE Dell Technologies

Media Relations, Media_Relations@Dell.com