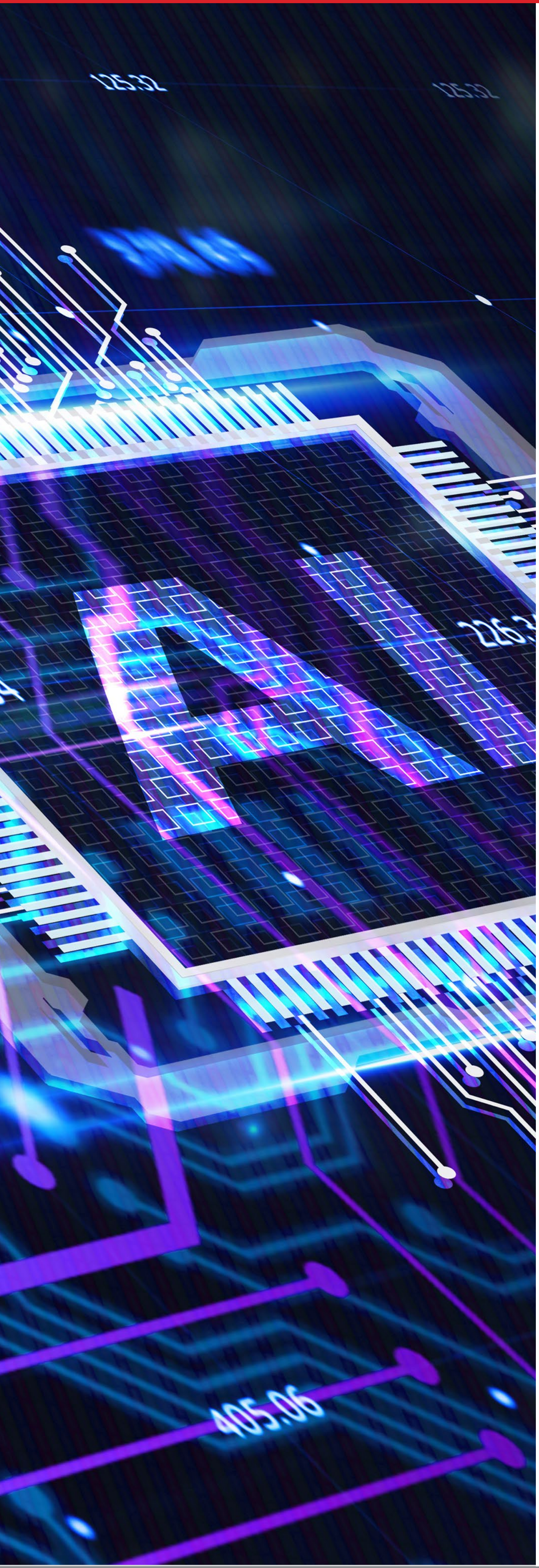


Meeting the challenges of AI workloads with the Dell AI portfolio

A comparison of the Dell AI portfolio vs. similar offerings from HPE

There's no doubt that artificial intelligence (AI) has transformed the business landscape and enabled industries and organizations of all sizes to gain deeper insights from their data, automate business processes, deliver tailored customer and user experiences, and better compete in their industry. To effectively harness the power of AI, organizations need an infrastructure provider that can offer a comprehensive and integrated solution portfolio that encompasses the entire AI lifecycle.

To help customers address the growing demands of AI and navigate its inherent complexities are infrastructure vendors such as Dell Technologies and Hewlett Packard Enterprise (HPE). Showcasing AI-ready portfolios, these vendors offer differing levels of AI solutions that unite high-performance on-prem and cloud infrastructure solutions with strategic partnerships and a menu of support and consultation services.

This report examines publicly available information about the Dell and HPE AI portfolios* with a goal of highlighting specific architectural, performance, and support advantages that customers might benefit from by selecting Dell Technologies for their AI needs. We compare details of the servers that Dell built to support AI deployments and reference industry benchmark testing results from ML Commons®. We also explore additional software and service offerings that support customers at every stage of their AI journey.

*Note: PT completed all research on or before December 5, 2023, so this paper will not reflect offerings or changes either Dell or HPE releases after that date.

The challenges of adopting AI

Adopting an AI strategy presents many new challenges for data centers and the IT personnel who staff them including:

- Addressing the existing skill gaps in their current staff through either in-house AI training or external hiring.
- Understanding the data preparation needs of AI, including the quality, quantity, location, and current state of the business's data.
- Assessing the specific business AI goals to better determine which AI models and implementations will provide benefits.
- Assessing the computational, networking, and storage needs of the planned AI systems, and determining an acquisition plan.

These are just a few examples of the many, often significant hurdles a company will face when looking to reap the benefits of implementing AI in their data centers.

The Dell AI portfolio seeks to help customers address these challenges through professional and consultative services that help customers build implementation roadmaps and prepare their data for AI models.[1] The portfolio also includes training courses that cover machine learning (ML) concepts and other educational topics and offers validated designs for AI to help ensure implementation success.[2] In addition, Dell partners with third parties to bring to customers additional AI tools , such as a custom Dell portal within the Hugging Face community with dedicated containers and scrips for open-source AI model deployment[3] and easy deployment of the Meta Llama 2 large language model (LLM).[4] Along with a large selection of compute and PC offerings, from mobile workstations to servers that support up to 8 high-end NVIDIA GPUs, Dell also provides the unstructured data storage AI requires with a portfolio of high-performance file and object storage arrays. These storage offerings, including Dell PowerScale, ObjectScale, ECS, and onboard storage, can handle the unstructured data that AI workloads frequently employ.[5] Dell has also partnered with Snowflake to provide a hybrid cloud storage solution for Dell customers.[6] According to Dell analysis, as of August 2023, they offer the "broadest Generative AI Portfolio," going beyond just servers and storage by providing resources across the AI implementation journey.[7]

AI performance and accelerated compute options: Dell vs. HPE

AI workloads can use CPUs, GPUs, or both as computational resources depending on the size or type of workload. Some CPUs provide AI-specific accelerators, such as Intel Advanced Matrix Extensions (Intel AMX) in the latest Intel Xeon Scalable processors.[8] GPUs are often better for larger and/or more demanding workloads, but GPU form factor can affect performance levels. For example, some NVIDIA A100 and H100 model GPUs come in either universal PCIe or proprietary SXM form factors, the latter of which use the higher-performing NVIDIA SXM architecture.[9] Large memory capacities and server design features such as cooling architecture and power efficiency also affect performance. Most data centers still use air cooling, which means that high-performing compute (HPC) workloads need servers built to cool with air as effectively as possible. Below, we highlight PowerEdge server offerings in terms of components, cooling options, and more, along with their published MLCommons® MLPerf® scores.

AI model benchmark performance: MLPerf result comparison

MLPerf® is a benchmark suite that tests AI performance for both training and inferencing. For an organization to publish official MLPerf® results, the results must be compliant with specific conditions set by the benchmark developer, MLCommons®.[10] These compliance guidelines provide standards that make it easier to compare performance. For inference testing, MLPerf® uses Datacenter, Edge, Mobile, and Tiny datasets, and reports AI scores and watts of power consumed during testing. The inference benchmark suite includes testing for many common AI, ML, and DL models; see Table 1.

Table 1: AI, ML, and DL models included in MLPerf® testing and typical use cases for each. Source: Principled Technologies.

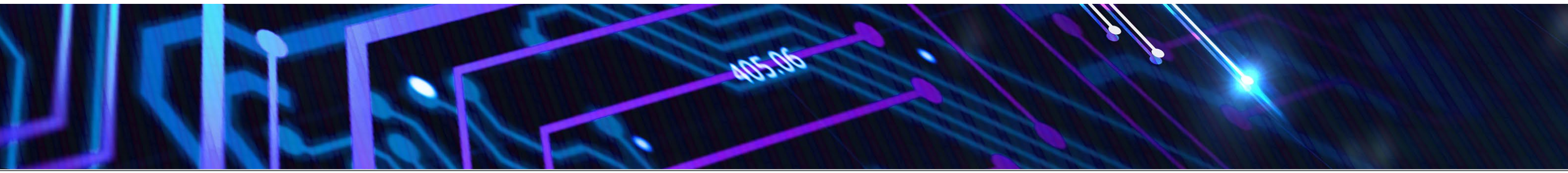
Common AI models	Typical use cases
ResNet	An image classification model that helps computers learn, remember, and identify different images for use cases such as medical imaging, social media content moderation, and facial recognition
RetinaNet	A type of object detection that can handle additional complexity compared to ResNet. It helps computers to identify and locate objects within images or video frames, and can classify them by importance. Used for things like autonomous driving, vehicle auto-assist technology, surveillance, facial recognition
3D-UNet	Specific to medical image segmentation
RNN-T	Speech recognition for use cases such as automated language translation
BERT	Natural language processing for use cases such as text summarization, language translation, and autocompletion of tasks
DLRM-v2-99.9	Recommendation model for use cases such as targeted ads and personalized product recommendations
GPTJ-99 and 99.9	LLM for natural language processing that excels at text generation for use cases such as chatbots and chat-based AI tools

MLPerf

MLPerf® results include several parameters in addition to the AI models themselves, which can make for a lot of data to parse in a single chart or table. Here's a quick reference to these parameters:

- 99.0 and 99.9: These numbers refer to the accuracy to which the model was trained. The more accurate you need the output to be, the more complex the model and the longer it can take to process data.
- Offline samples/sec: Mode where the benchmark sends all queries at the beginning of the test simulating data already present on the system.
- Server queries/sec: Mode where the benchmark sends queries throughout the test duration simulating analyzing a live stream of data.

For more about MLCommons® and MLPerf® results, see https://mlcommons.org/benchmarks/inference-datacenter/.



Results in this report come from MLPerf® v3.1 Inference Datacenter results published on the MLCommons® website from November 2023.[11] These results include submissions from technology manufacturers and cloud service providers and cover a range of configurations. Compared to publicly available submissions from HPE, Dell servers produced better results in certain AI models. (Note: Different GPU configurations across the servers can make head-to-head comparisons difficult.) See Table 2 for details.

Table 2: Dell and HPE servers included in the MLCommons® MLPerf® 3.1 results published as of 11/29/23.
Source: Principled Technologies.

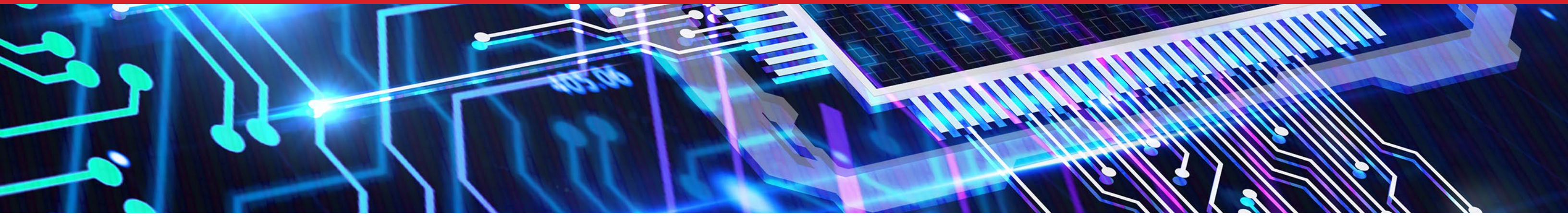
Submitter	Server model	# and model of GPUs	Description
Dell[12]	PowerEdge XE9680	8x NVIDIA H100 SXM	For AI training and inference with large workloads such as large language models
	PowerEdge XE9640	4x NVIDIA H100 SXM	For training large AI models in high density and liquid cooled datacenters
	PowerEdge XE8640	4x NVIDIA H100 SXM	For driving traditional AI training, HPC, and data analytics apps in a 4U form factor for air-cooled datacenters
	PowerEdge R760xa	4x NVIDIA H100 PCIe	For a wide range of high-compute workloads including AI-ML/DL training and inferencing that do not require top performing GPUs
HPE[13,14]	ProLiant XL675d Gen10 Plus	8x NVIDIA A100 SXM	For high-performance computing and AI
	ProLiant DL380a Gen11	4x NVIDIA H100 PCIe	2U server for moderate AI workloads

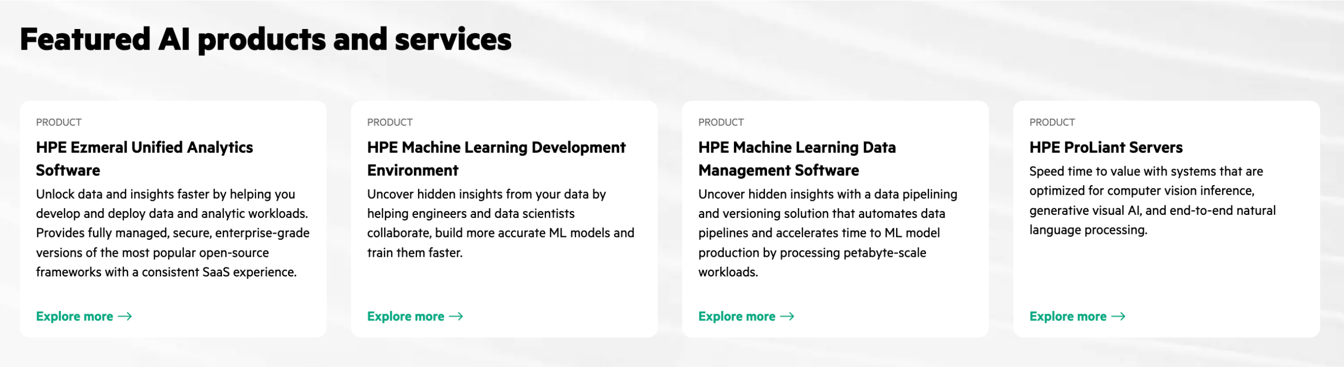
Direct comparison between Dell and HPE servers

While there is more to a comprehensive AI strategy than hardware, ensuring the strongest hardware performance is one of the most vital factors in AI workload success. As newer GPUs and other technologies become available, AI workload capabilities also evolve. At the time the MLPerf® v3.1 results were first published, the best NVIDIA GPU available was the H100 Tensor Core with which Dell published MLPerf® results in several of their servers in both PCIe and SXM5 form factors.[15] The published HPE results included only one H100 submission and with only the PCIe form factor. Our research showed that few of the available GPU-capable HPE servers supported the SXM5 H100 form factor for the best NVIDIA GPU performance, and none of the HPE ProLiant servers do.[16] As we show below, having better GPUs typically improves AI workload performance.

Eight-GPU MLPerf results

The Dell PowerEdge XE9680 offers support for up to eight NVIDIA H100 SXM5 GPUs for AI acceleration and up to two 4th Generation Intel® Xeon® Scalable processors. The PowerEdge XE product family has a modular architecture supporting SXM4 or SXM5 NVIDIA GPUs or Open Compute Project Accelerator Module (OAM) GPU assemblies from AMD, which can boost performance compared to a standard PCIe GPU.[17] Taking up only 6U of rack space, the PowerEdge XE9680 is a compact eight-way NVIDIA H100 SXM5 server. The latest HPE ProLiant Gen11 servers do not currently support the H100 SXM form factor,[18] though some of the HPE Cray Supercomputing servers do.[19] Because HPE did not submit any MLPerf® results with the Cray servers and highlights only their ProLiant servers on their AI portfolio page, we will focus on ProLiant servers for this paper. (See Figure 1.)





Featured AI products and services

PRODUCT

HPE Ezmeral Unified Analytics Software

Unlock data and insights faster by helping you develop and deploy data and analytic workloads. Provides fully managed, secure, enterprise-grade versions of the most popular open-source frameworks with a consistent SaaS experience.

Explore more →

PRODUCT

HPE Machine Learning Development Environment

Uncover hidden insights from your data by helping engineers and data scientists collaborate, build more accurate ML models and train them faster.

Explore more →

PRODUCT

HPE Machine Learning Data Management Software

Uncover hidden insights with a data pipelining and versioning solution that automates data pipelines and accelerates time to ML model production by processing petabyte-scale workloads.

Explore more →

PRODUCT

HPE ProLiant Servers

Speed time to value with systems that are optimized for computer vision inference, generative visual AI, and end-to-end natural language processing.

Explore more →

Figure 1: Screenshot of Featured AI products and services at https://www.hpe.com/us/en/solutions/ai-artificial-intelligence.html highlighting HPE ProLiant Servers as of 12/5/2023.

In the MLPerf® v3.1 results first published in November 2023 for eight-GPU servers, the Dell PowerEdge XE9680 with NVIDIA SXM5 H100 GPUs outperformed the HPE ProLiant XL675d Gen10 Plus with NVIDIA SXM4 A100 GPUs by up to 4.25x (see Figure 2).

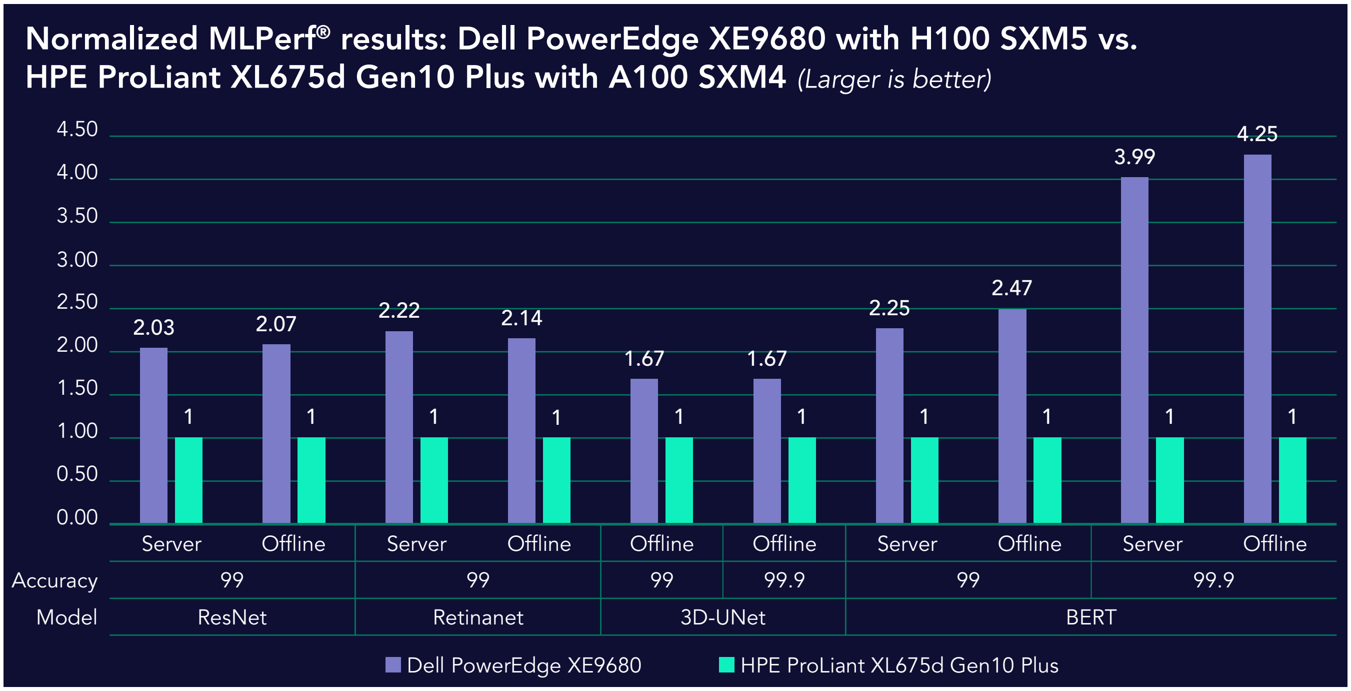


Figure 2: Published MLPerf® results for the Dell PowerEdge XE9680 and HPE ProLiant XL675d Gen10 Plus as of 11/29/23. The Dell System uses the NVIDIA H100 GPU, while the GPUs in the HPE system are one generation older. Source: Principled Technologies using data from MLCommons®.[20,21]

For ease of comparison, we have normalized the test results in Figures 2 through 5. This means we assign the value of 1 to each HPE ProLiant DL380a Gen 11 result and show the corresponding Dell PowerEdge R760xa result in relation to it. As these results show, even one generation difference between GPU models can make a significant difference in the performance you can expect to see across a multitude of AI workloads.

Four-GPU MLPerf results

When saving data center power or space savings are key concerns, the 2U Dell PowerEdge XE9640 could provide the answer. With up to four NVIDIA H100 SXM GPUs, the PowerEdgeXE9640 offers half the GPU computational power of the XE9680, in two-thirds less space.[22] The densely packed Dell PowerEdge XE9640 incorporates Dell Smart Cooling technology, providing an array of thermal technology including direct liquid cooling for CPUs and GPUs.[23]

The 2U chassis of the PowerEdge XE9640 accommodates improved airflow mechanisms, including larger fans and heatsinks, to help cool the other vital components, such as PCIe cards and memory.[24] The PowerEdge XE9640 is currently the only offering from either Dell or HPE that comes with four-way HGX H100 GPUs at 2U. The HPE AI portfolio offers 1U and 2U Gen11 ProLiant servers, but they are limited to PCIe form factor GPUs.[25]

The Dell PowerEdge XE9640 server also supports Intel Max GPU Series 1550 OAM GPUs, which provide a low-power, high-density GPU option that includes a PCIe card and an OpenCompute Accelerator Module (OAM).[26] While we could not ascertain that as of 12/5/23 HPE offered a server with these GPUs, they do offer the HPE ProLiant DL380 Gen11 and DL380a Gen11 with Intel Data Center GPU Max 1100 GPUs.[27] This means that the Dell PowerEdge XE9640 could be the only current offering with four Intel Max 1550 OAM GPUs in a 2U server. For companies concerned with datacenter space and power efficiency, a 2U server with four Intel Max 1550 GPUs provides a solution that marries high-performance compute and energy efficiency without sacrificing data center space.

The Dell PowerEdge XE9640 with four HGX H100 GPUs outperformed the HPE ProLiant DL380a with four PCIe H100 GPUs by up to 1.99x in the published MLPerf® 3.1 results (see Figure 3).

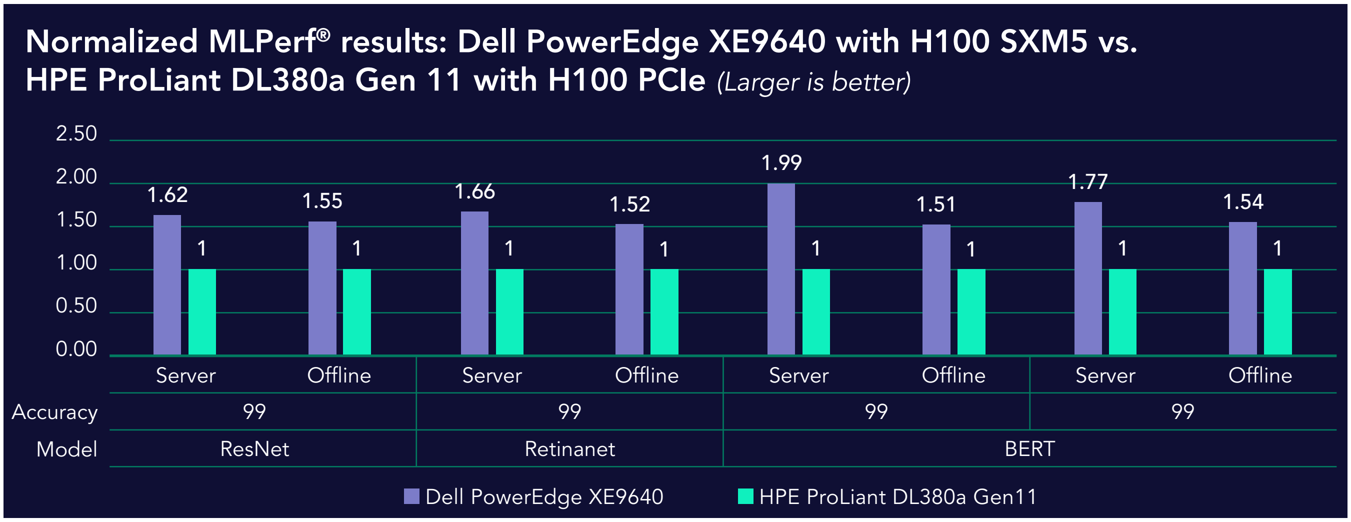
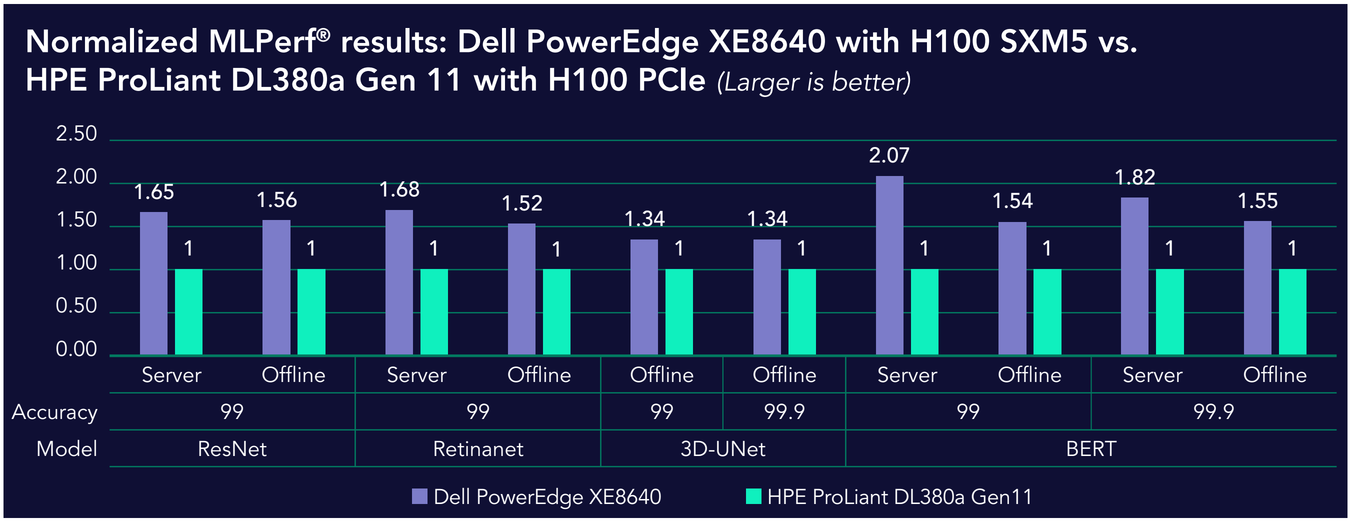

Normalized MLPerf® results: Dell PowerEdge XE9640 with H100 SXM5 vs. HPE ProLiant DL380a Gen 11 with H100 PCIe *(Larger is better)*

Figure 3: Published MLPerf® results for the Dell PowerEdge XE9680 and HPE ProLiant XL675d Gen10 Plus as of 11/29/23. The Dell System uses the NVIDIA H100 GPU, while the GPUs in the HPE system are one generation older. Source: Principled Technologies using data from MLCommons®.[28,29]

The PowerEdge XE8640 offers a four-way GPU configuration with air cooling for processors and a Liquid Assisted Air Cooling Radiator for the GPUs which does not require facility water –to-rack availability. For those who do not or cannot use external coolant,[30] the 4U Dell PowerEdge XE8640 supports four NVIDIA H100 SXM5 GPUs providing the same computational power as the PowerEdge XE9640 without the need for direct liquid cooling.[31]

The Dell PowerEdge XE8640 features the latest 4th Generation Intel Xeon Scalable processors and up to 4 TB of memory[32] to handle the large datasets and complex computations common in AI and data analytics. Again, HPE does offer the NVIDIA H100 SXM5 GPUs in HPE Cray systems, but HPE ProLiant GPU-enabled servers do not support it.

Compared to MLPerf® data published as of November 2023, the PowerEdge XE8640 server with four NVIDIA H100 SXM5 GPUs achieved the highest AI throughput among all four-GPU submissions in nine different categories. As Figure 4 shows, compared to the HPE ProLiant DL380a server, it scored up to 2.07 times as high.



Normalized MLPerf® results: Dell PowerEdge XE8640 with H100 SXM5 vs. HPE ProLiant DL380a Gen 11 with H100 PCIe *(Larger is better)*

Figure 4: Published MLPerf® results for the Dell PowerEdge XE8640 and HPE ProLiant DL380a Gen11 as of 11/29/23. The Dell system uses the NVIDIA H100 SXM form factor, while the HPE system uses the less powerful PCIe form factor. Source: Principled Technologies using data from MLCommons®.[33,34]

Finally, for organizations who may wish to start smaller and grow as needed, the 2U Dell PowerEdge R760xa server accommodates a range of GPUs from NVIDIA, AMD, and Intel, with support for up to four double-width PCIe Gen 5 GPUs or 12 single-width PCIe GPUs.[35] It features 32 DIMM slots, an eight-drive bay for 2.5-inch disks, and 12 PCIe slots, providing scalable storage that can grow with increasing AI data requirements and support for up to 12 single-width PCIe GPUs or four double-width PCIe GPUs like the NVIDIA H100 or L40S.[36] This scalability means that the server can adapt to evolving AI tasks, from machine learning model training to advanced data processing.

The air-cooling system of the PowerEdge R760xa supports high-density computing environments and can accommodate higher thermal design power (TDP) accelerators up to 350W,[37] an ability that can help it maintain performance under intensive computational loads. In MLPerf® ResNet, RetinaNet and BERT Server test results published as of November 2023 using the "server" mode, the PowerEdge R760xa with four NVIDIA H100 PCIe GPUs outperformed the HPE ProLiant DL380a Gen 11 also equipped with four H100 PCIe GPUs (see Figure 5).

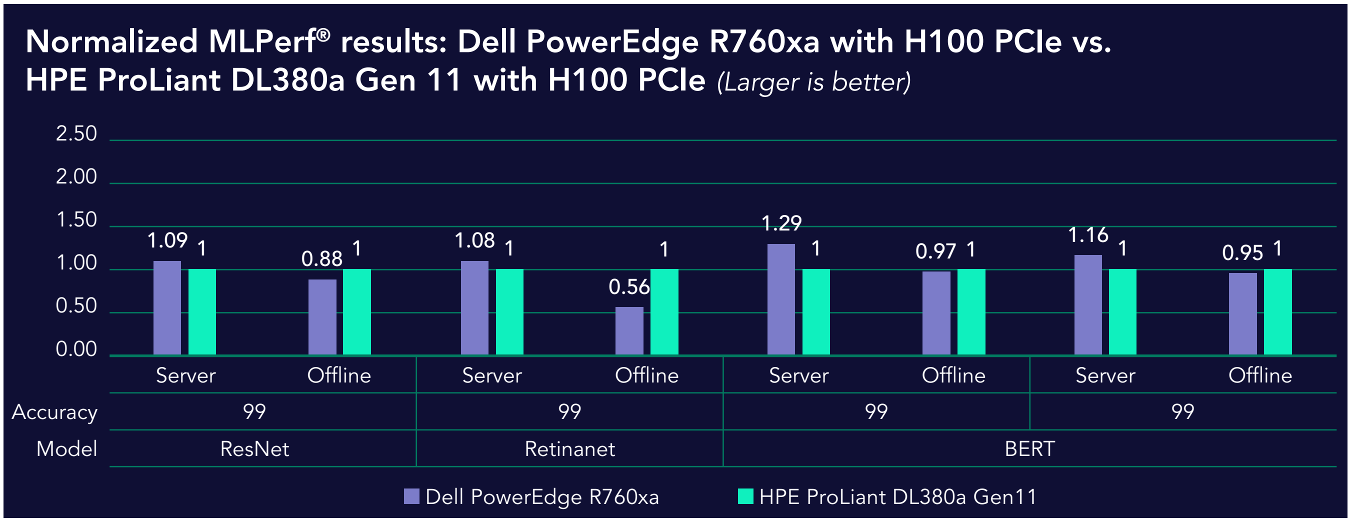


Figure 5: Published MLPerf® results for the Dell PowerEdge R760xa and HPE ProLiant DL380a Gen11 as of 11/29/23. Both systems use the PCIe form factor of the NVIDIA H100 GPUs. Source: Principled Technologies using data from MLCommons®.[38,39]

All in all, the MLPerf® results show that performance varies widely across servers and components, so selecting the right options to support your workloads and their performance demands is critical. Dell PowerEdge servers for AI workloads offer multiple options for cooling and density to fit whatever data center needs a company may have while providing strong MLPerf® performance.

More detailed coverage of the Dell AI Portfolio

While crucial, compute performance is only one consideration when planning your AI workloads. You must also consider the rest of the AI portfolio a vendor offers when embarking upon an AI implementation. Below we discuss additional categories critical to these AI portfolios, including client workstations, cloud-native products, storage, and more. We also highlight areas where offerings from Dell may offer an advantage compared to HPE.

Workstations

For AI developers and data scientists, Dell Precision Data Science workstations offer NVIDIA RTX™ GPUs and Intel Xeon® CPUs, along with a suite of data science tools.[40] These systems leverage professional-grade compute options with NVIDIA GPUs certified for over 100 professional applications[41] and Intel Xeon Scalable processor accelerators such as Intel DL Boost.[42] Precision workstations come in mobile, tower, and rack formats to serve needs ranging from larger, stationary data analysis to on-the-go scientific field modeling.

HPE workstation offerings are narrower, primarily featuring single workstation towers equipped with NVIDIA L4s; HPE offers no mobile workstation option.[43] While adequate for many tasks, their workstation tower offerings don't bring the same flexibility and workload coverage as the more extensive range Dell provides. The variety in size and portability options in Dell Precision workstations allow for more tailored solutions, accommodating the different needs in settings such as laboratories, offices, and field operations.

Storage

Storage may be just as vital as compute to running AI workloads. More data improves AI model accuracy, but storing and managing massive datasets can challenge many data centers' capabilities. Additionally, because models are typically trained using unstructured data, AI-ready storage systems must handle many different data types with ease.[44] To provide capacity and scaling for AI, ML, and DL datasets, Dell offers the PowerScale™ series for file storage and Elastic Cloud Storage (ECS) or software-defined ObjectScale for object storage.

The PowerScale all-flash NAS portfolio offers capacity options ranging from3.84TB up to 720TB raw capacity per node, with clustered all-flash capacities reaching 186PB of raw capacity. The flexibility and scale of PowerScale can support a wide variety of customers and AI use cases.[45] When clustered, the PowerScale F900 can reach up to 186PB of total raw storage.[46] All three all-flash PowerScale models—F200, F600, and F900—include inline data compression and deduplication to improve storage efficiency.[47] Each PowerScale storage model uses the Dell OneFS™ file system, which employs policies to tier storage to prioritize the most important data on the fastest tiers for workload optimization.[48] Dell also offers OneFS software in the AWS marketplace with Dell APEX File Storage for AWS. Customers can leverage OneFS with their AWS compute instances for a consistent user experience with the same features available in on-premises OneFS arrays.[49] While HPE does offer public cloud integration for hybrid storage solutions, we did not find a cloud-native option like Dell APEX File Storage for AWS among its offerings.

Object storage options from Dell include Dell Enterprise Object Storage (ECS), which is "purpose-built to store unstructured data at public cloud scale."[50] Along with built-in compatibility with Amazon S3 object storage for hybrid cloud functionality, ECS storage nodes deliver capacities up to 14PB per rack.[51] HPE also offers unstructured storage with file and object storage options, though its object storage offering is via a partnership with Scality. Customers can purchase HPE Solutions for Scality from HPE.[52]

Professional services

Dell offers a wide spectrum of professional services, including consulting, data preparation, deployment, support, and managed services to support AI deployments. For organizations looking for validated architectures and solutions, Dell offers Validated Designs for AI, which target specific use cases to take the guesswork out of designing and deploying AI resources. These Dell-validated AI solutions include hardware and software bundles, conversational AI models, machine learning operations, and more. By combining pre-configured, purpose-built solutions with AI-related services, Dell offers a comprehensive AI solution across the spectrum of AI needs. These offerings could provide a quicker and easier path to AI success compared to building ad-hoc solutions.

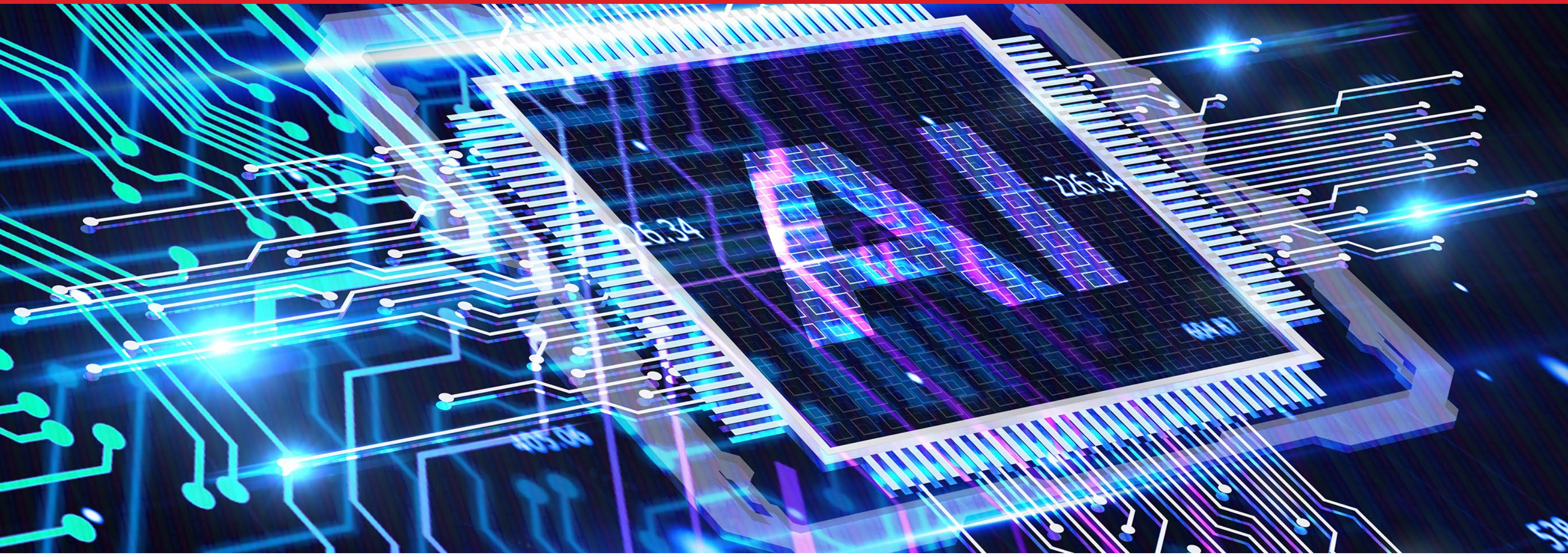
Dell services can also guide your AI journey from advising to implementation. Dell ProConsult Advisory Services helps customers identify where users can benefit from adopting GenAI processes and create a roadmap comprising required solutions and IT skills. Dell services can prepare data for Large Language Model integration and train IT teams on AI knowledge. For full GenAI adoption, Dell teams review your specific use cases and determine, deploy, and configure the best AI model to suit your needs. HPE also offers professional services to support companies in their AI endeavors.[53,54]

Management considerations

Servers require ongoing management that takes up admin time. Firmware, software, and drivers need periodic updates, IT staff must optimize and maintain performance and temperatures, and more. In previous Principled Technologies (PT) testing, we assessed the management capabilities of Dell servers with Integrated Dell Remote Access Controller 9 (iDRAC9).[55] Admins can rely on automated online updates with iDRAC9 OpenManage™ Enterprise (OME) with configurable scheduling to keep their servers up-to-date and use profiles to quickly and easily onboard new servers as workloads grow. With iDRAC and OME, Dell customers can access more remote management features, deploy servers easier, and update firmware easier than they could using HPE OneView and HPE iLO. With Dell PowerEdge servers come Dell management and services that could help organizations by "reducing time and effort for tasks like monitoring system health or updating firmware," freeing up IT cycles for innovation and other tasks.[56]

Table 3: Summary of comparison between Dell and HPE management tools from a November 2022 PT report.[57]
Source: Principled Technologies.

	What's different with Dell management tools	How much better
More remote management features iDRAC vs. iLO	More HTML5 console and BIOS configuration features for more remote functionality in iDRAC	2.5X the HTML5 console features and 13X the BIOS features
Easier server deployment OME vs. OneView	One-to-many profile deployment with OME	52% less time to deploy a server than with OneView
Easier firmware updates OME vs. OneView	Automated online updates with OME	Update multiple servers by connecting to Dell.com, saving the time it takes to update servers by manually uploading bundles with OneView
Easier alerting OME vs. OneView	Set up alert policies in OME and execute automated actions based on alerts	Automating this process saves time and reduces potential for errors vs. executing actions manually each time you receive an alert in OneView
Easier to use security features (system lockdown and dynamic USB) iDRAC vs. iLO	Fewer steps, less time, no reboots using iDRAC	¼ steps, 91% less time for System Lockdown
More robust analytics CloudIQ for PowerEdge vs. InfoSight	Customizable reports, more health metrics for better admin control with CloudIQ for PowerEdge	Over 15x more metrics to choose from compared to InfoSight



Conclusion

Harnessing the power of AI to streamline and improve business operations can be a challenging task, with significant business implications. With technology advancing more rapidly than ever, partnering with the right vendor for AI is key. By choosing a company like Dell that not only offers a comprehensive AI portfolio, but can also provide planning, preparation, implementation and management services, customers can face these challenges head on. MLPerf® Benchmark testing shows that offerings in the Dell AI portfolio offer consistent, strong performance for AI workloads. With high-performing and flexible server options, along with multiple storage choices, validated solutions, and professional services specifically tailored for AI, Dell can help businesses embrace AI and its benefits.

[1] Dell, "Increasing Your Data Value with Dell Generative AI Solutions," accessed December 19, 2023, https://www.dell.com/en-us/blog/increasing-your-data-value-with-dell-generative-ai-solutions/.

[2] Dell, "Dell AI solutions," accessed December 12, 2023, https://www.dell.com/en-us/dt/solutions/artificial-intelligence/index.htm#accordion0&tab0=0.

[3] Dell, "Dell Technologies and Hugging Face to Simplify Generative AI with On-Premises IT," accessed December 12, 2023, https://www.dell.com/en-us/dt/corporate/newsroom/announcements/detailpage.press-releases~usa~2023~11~20231114-dell-technologies-and-hugging-face-to-simplify-generative-ai-with-on-premises-it.htm#/filter-on/Country:en-us.

[4] Dell, "Dell and Meta Collaborate to Drive Generative AI Innovation," accessed December 12, 2023, https://www.dell.com/en-us/blog/dell-and-meta-collaborate-to-drive-generative-ai-innovation/.

[5] Dell, "Dell AI-Ready Data Platform," accessed December 12, 2023, https://www.dell.com/en-us/dt/solutions/artificial-intelligence/storage-for-ai.htm?hve=explore+unstructured+storage#tab0=0.

[6] Dell, "Snowflake and Dell Partnership Gains Momentum," accessed December 19, 2023, https://www.dell.com/en-us/blog/snowflake-and-dell-partnership-gains-momentum/.

[7] Robert McNeal, "Dell, VMware and NVIDIA Bring AI to Your Data," accessed January 17, 2024, https://www.dell.com/en-us/blog/dell-vmware-and-nvidia-bring-ai-to-your-data/. Per the link above: "Based on Dell analysis, August 2023. Dell Technologies offers solutions engineered to support AI workloads from Workstations PCs (mobile and fixed) to Servers for High-performance Computing, Data Storage, Cloud Native Software-Defined Infrastructure, Networking Switches, Data Protection, HCI and Services."

[8] Intel, "Accelerate Artificial Intelligence (AI) Workloads with Intel Advanced Matrix Extensions (Intel AMX)," accessed December 12, 2023, https://www.intel.com/content/www/us/en/content-details/785250/accelerate-artificial-intelligence-ai-workloads-with-intel-advanced-matrix-extensions-intel-amx.html.

[9] Vipera, "NVIDIA's H100 and A100 GPU Cards: Exploring the Intricacies of SXM and PCI-E Connections," accessed December 12, 2023, https://www.viperatech.com/unraveling-the-mysteries-sxm-vs-pci-e-connections-in-nvidias-high-end-h100-and-a100-gpus/.

[10] GitHub, "MLPerf® Results Messaging Guidelines," accessed January 16, 2024, https://github.com/mlcommons/policies/blob/master/MLPerf_Results_Messaging_Guidelines.adoc.

[11] MLCommons®, "MLPerf® Inference: Datacenter Benchmark Suite Results," accessed December 12, 2023, https://mlcommons.org/en/inference-datacenter-31/.

[12] Dell, "PowerEdge XE Servers," accessed December 12, 2023, https://www.dell.com/en-us/dt/servers/specialty-servers/poweredge-xe-servers.htm?hve=explore+poweredge+xe#tab0=0.

[13] HPE, "HPE ProLiant XL675d Gen10 Plus Configure-to-order Server," accessed December 12, 2023, https://www.hpe.com/us/en/product-catalog/compute/proliant-servers/pip.1013142988.html.

[14] HPE, "HPE ProLiant DL380a Gen11," accessed December 12, 2023, https://www.hpe.com/us/en/product-catalog/compute/proliant-servers/pip.proliant-dl380-server.1014696168.html.

[15] MLCommons®, "MLPerf® Inference: Datacenter Benchmark Suite Results v 3.1," accessed December 12, 2023, https://mlcommons.org/benchmarks/inference-datacenter/.

[16] HPE, "NVIDIA Accelerators for HPE ProLiant Servers," accessed December 12, 2023, https://www.hpe.com/psnow/doc/c04123180.html?jumpid=in_pdp-psnow-qs.

[17] Dell, "PowerEdge XE9680 Specification Sheet," accessed January 19, 2024, https://www.delltechnologies.com/asset/en-us/products/servers/technical-support/poweredge-xe9680-spec-sheet.pdf.

[18] HPE, "HPE & NVIDIA financial services solution sets new records in performance," accessed December 12, 2023, https://community.hpe.com/t5/alliances/hpe-amp-nvidia-financial-services-solution-sets-new-records-in/ba-p/7197388.

[19] HPE, "QuickSpecs: HPE Cray Supercomputing XD670," accessed December 12, 2023, https://www.hpe.com/psnow/doc/a50004292enw.

[20] Verified MLPerf® score of v3.1 Inference Closed. Retrieved from https://mlcommons.org/benchmarks/inference-datacenter/ 5 December 2023, entry 3.1-0069. The MLPerf® name and logo are registered and unregistered trademarks of ML-Commons® Association in the United States and other countries. All rights reserved. Unauthorized use strictly prohibited. See www.mlcommons.org for more information.

[21] Verified MLPerf® score of v3.1 Inference Closed. Retrieved from https://mlcommons.org/benchmarks/inference-datacenter/ 5 December 2023, entry 3.1-0085. The MLPerf® name and logo are registered and unregistered trademarks of ML-Commons® Association in the United States and other countries. All rights reserved. Unauthorized use strictly prohibited. See www.mlcommons.org for more information.

[22] Dell, "PowerEdge XE9640 Rack Server," accessed December 12, 2023, https://www.dell.com/en-us/shop/ipovw/poweredge-xe9640.

[23] Accelsius, "Enabling the AI Revolution with Liquid Cooling," accessed December 12, 2023, https://www.accelsius.com/blog/enabling-the-ai-revolution-with-liquid-cooling.

[24] Dell, "Dell PowerEdge XE9640 Technical Guide," accessed December 12, 2023, https://www.delltechnologies.com/asset/en-us/products/servers/technical-support/poweredge-xe9640-technical-guide.pdf.

[25] HPE, "HPE ProLiant DL380a Gen11," accessed December 12, 2023, https://www.hpe.com/psnow/doc/PSN1014696168WWEN.pdf?jumpid=in_pdp-psnow-dds.

[26] Intel, "Intel® Data Center GPU Max Series Technical Overview," accessed December 12, 2023, https://www.intel.com/content/www/us/en/developer/articles/technical/intel-data-center-gpu-max-series-overview.html#gs.08874l.

[27] HPE, "Intel Data Center GPU Max 1100 48GB Accelerator for HPE Data sheet," accessed December 12, 2023, https://www.hpe.com/psnow/doc/PSN1014779728WWEN.

[28] Verified MLPerf® score of v3.1 Inference Closed. Retrieved from https://mlcommons.org/benchmarks/inference-datacenter/ 5 December 2023, entry 3.1-0066. The MLPerf® name and logo are registered and unregistered trademarks of ML-Commons® Association in the United States and other countries. All rights reserved. Unauthorized use strictly prohibited. See www.mlcommons.org for more information.

[29] Verified MLPerf® score of v3.1 Inference Closed. Retrieved from https://mlcommons.org/benchmarks/inference-datacenter/ 5 December 2023, entry 3.1-0084. The MLPerf® name and logo are registered and unregistered trademarks of ML-Commons® Association in the United States and other countries. All rights reserved. Unauthorized use strictly prohibited. See www.mlcommons.org for more information.

[30] Dell, “AI and HPC —With Air or Liquid Cooling,” accessed December 12, 2023, https://www.delltechnologies.com/asset/en-us/products/servers/briefs-summaries/poweredge-xe9640-and-xe8640-infographic.pdf.

[31] Dell, “PowerEdge XE8640 : Drive AI, HPC modeling and simulation workloads with superior performance,” accessed December 12, 2023, https://www.delltechnologies.com/asset/en-us/products/servers/technical-support/poweredge-xe8640-spec-sheet.pdf.

[32] Dell, “PowerEdge XE8640 Rack Server,” accessed December 12, 2023, https://www.dell.com/en-us/shop/ipovw/poweredge-xe8640.

[33] Verified MLPerf® score of v3.1 Inference Closed. Retrieved from https://mlcommons.org/benchmarks/inference-datacenter/ 5 December 2023, entry 3.1-0067. The MLPerf® name and logo are registered and unregistered trademarks of MLCommons® Association in the United States and other countries. All rights reserved. Unauthorized use strictly prohibited. See www.mlcommons.org for more information.

[34] Verified MLPerf® score of v3.1 Inference Closed. Retrieved from https://mlcommons.org/benchmarks/inference-datacenter/ 5 December 2023, entry 3.1-0084. The MLPerf® name and logo are registered and unregistered trademarks of MLCommons® Association in the United States and other countries. All rights reserved. Unauthorized use strictly prohibited. See www.mlcommons.org for more information.

[35] Dell, “PowerEdge R760xa Rack Server,” accessed December 12, 2023, https://www.dell.com/en-us/shop/dell-poweredge-servers/poweredge-r760xa-rack-server/spd/poweredge-r760xa/pe_r760xa_16902_vi_vp#features_section.

[36] SANStorageWorks, “Dell EMC PowerEdge R760xa: Powerful and scalable for GPU workloads,” accessed December 12, 2023, https://www.sanstorageworks.com/PowerEdge-R760xa.asp.

[37] Dell, “Dell PowerEdge Servers and NVIDIA GPUs,” accessed December 12, 2023, https://infohub.delltechnologies.com/l/design-guide-generative-ai-in-the-enterprise-inferencing/dell-poweredge-servers-and-nvidia-gpus-1/.

[38] Verified MLPerf® score of v3.1 Inference Closed. Retrieved from https://mlcommons.org/benchmarks/inference-datacenter/ 5 December 2023, entry 3.1-0064. The MLPerf® name and logo are registered and unregistered trademarks of MLCommons® Association in the United States and other countries. All rights reserved. Unauthorized use strictly prohibited. See www.mlcommons.org for more information.

[39] Verified MLPerf® score of v3.1 Inference Closed. Retrieved from https://mlcommons.org/benchmarks/inference-datacenter/ 5 December 2023, entry 3.1-0084. The MLPerf® name and logo are registered and unregistered trademarks of MLCommons® Association in the United States and other countries. All rights reserved. Unauthorized use strictly prohibited. See www.mlcommons.org for more information.

[40] Dell, “Workstations for AI,” accessed December 12, 2023, https://www.dell.com/en-us/dt/ai-technologies/index.htm?hve=explore+dell+precision+for+ai#pdf-overlay=//www.delltechnologies.com/asset/en-us/products/workstations/briefs-summaries/ai-industry-brochure.pdf.

[41] NVIDIA, “NVIDIA RTX in Professional Workstations,” accessed December 12, 2023, https://www.nvidia.com/en-us/design-visualization/desktop-graphics/.

[42] Intel, “Intel® Deep Learning Boost (Intel® DL Boost),” accessed December 12, 2023, https://www.intel.com/content/www/us/en/artificial-intelligence/deep-learning-boost.html.

[43] HPE, “HPE ProLiant ML350 Gen11,” accessed December 12, 2023, https://buy.hpe.com/us/en/compute/tower-servers/proliant-ml300-servers/proliant-ml350-server/hpe-proliant-ml350-gen11/p/1014696172.

[44] ComputerWeekly.com, “Storage requirements for AI, ML and analytics in 2022,” accessed December 12, 2023, https://www.computerweekly.com/feature/Storage-requirements-for-AI-ML-and-analytics-in-2022.

[45] Dell, “PowerScale AI-Ready Data Platform,” accessed December 12, 2023, https://www.dell.com/en-us/shop/powerscale-family/sf/powerscale.

[46] Dell, “Compare PowerScale,” accessed December 12, 2023, https://www.dell.com/en-us/shop/powerscale-family/sf/powerscale#compare-module.

[47] Dell, “Dell PowerScale All-Flash,” accessed December 12, 2023, https://www.delltechnologies.com/asset/en-us/products/storage/technical-support/h15963-ss-powerscale-all-flash-nodes.pdf.

[48] Dell, “Dell PowerScale OneFS Software Features,” accessed December 12, 2023, https://www.delltechnologies.com/asset/en-us/products/storage/technical-support/h18275-onefs-software-features-data-sheet.pdf.

[49] Dell, “Dell APEX File Storage for AWS,” accessed December 12, 2023, https://www.dell.com/en-us/dt/apex/storage/public-cloud/file.htm#pdf-overlay=//www.delltechnologies.com/asset/en-us/products/storage/briefs-summaries/h19575-so-apex-file-storage-for-aws.pdf.

50. Dell, “Dell ECS Enterprise Object Storage,” accessed December 12, 2023, https://www.dell.com/en-us/dt/storage/ecs/index.htm?hve=explore+ecs#tab0=0&tab1=0.

51. Dell, “Dell ECS Enterprise Object Storage,” accessed December 12, 2023, https://www.dell.com/en-us/dt/storage/ecs/index.htm#tab0=0&tab1=0&accordion0.

52. HPE, “Storage Solutions for Scality,” accessed December 12, 2023, https://www.hpe.com/us/en/storage/file-object/scality.html.

53. HPE, “Make AI work for you,” accessed January 16, 2024, https://www.hpe.com/us/en/solutions/ai-artificial-intelligence.html.

54. HPE, “HPE AI Services – Generative AI Implementation,” accessed January 16, 2024, https://www.hpe.com/us/en/services/generative-ai-implementation-service.html.

55. Principled Technologies, “Simplify administrator tasks and improve security and health monitoring with tools from the Dell management portfolio vs. comparable tools from HPE,” accessed December 12, 2023, https://www.principledtechnologies.com/Dell/Management-tools-vs-HPE-1122.pdf.

56. Principled Technologies, “Simplify administrator tasks and improve security and health monitoring with tools from the Dell management portfolio vs. comparable tools from HPE,” accessed December 12, 2023, https://www.principledtechnologies.com/Dell/Management-tools-vs-HPE-1122.pdf.

57. Principled Technologies, “Simplify administrator tasks and improve security and health monitoring with tools from the Dell management portfolio vs. comparable tools from HPE.”

This project was commissioned by Dell Technologies.

Principled Technologies®

Facts matter.®

DISCLAIMER OF WARRANTIES; LIMITATION OF LIABILITY:
Principled Technologies, Inc. has made reasonable efforts to ensure the accuracy and validity of its testing, however, Principled Technologies, Inc. specifically disclaims any warranty, expressed or implied, relating to the test results and analysis, their accuracy, completeness or quality, including any implied warranty of fitness for any particular purpose. All persons or entities relying on the results of any testing do so at their own risk, and agree that Principled Technologies, Inc., its employees and its subcontractors shall have no liability whatsoever from any claim of loss or damage on account of any alleged error or defect in any testing procedure or result.

In no event shall Principled Technologies, Inc. be liable for indirect, special, incidental, or consequential damages in connection with its testing, even if advised of the possibility of such damages. In no event shall Principled Technologies, Inc.’s liability, including for direct damages, exceed the amounts paid in connection with Principled Technologies, Inc.’s testing. Customer’s sole and exclusive remedies are as set forth herein.