Dell Validated Design for HPC Financial Services - Machine Learning

2023

DELL Technologies

The benefits—Dell Validated Designs simplify IT transformation, helping you solve challenges faster

Scale easily

- A flexible building block approach easily scales over time.
- Scale by adding resources such as memory or hard drives inside Dell PowerEdge servers.
- Add external storage with Dell PowerVault storage arrays or Dell PowerScale scale-out network-attached storage (NAS).
- Augment compute and or accelerated compute as needs increase

Reduce risk

- Dell Technologies engineers and industry experts work in collaboration with you and our partners to design, deploy and scale HPC solutions for specific applications. This saves time and reduces the risk of potential hardware and software issues.
- Around the world, more than 35,000 Dell Technologies Services experts are available every step of the way with consulting, education, deployment, management and support.
- With proven success in thousands of implementations worldwide, you can be confident growing with Dell Technologies.

Optimize investments

- Purpose-built HPC building blocks are tailored to speed deployment, help eliminate potential software and hardware issues, and optimize performance.
- Flexible, industry-standard building blocks of compute, networking and storage are tested and tuned with your HPC and AI applications by Dell Technologies engineering teams. Available consulting, education, deployment, support and remote management services optimize solution productivity and efficiency.

DELL Technologies

Dell Validated Design (DVD) for Risk Assessment

(Launched November 2022)

- Scalable Rack Units for compute and accelerated compute serve as baselines to architect scalable custom clusters
- Combine Scalable Rack Units to support compute-centric workloads, data-centric workloads, or both with the same cluster
- Validated by Dell with benchmarking for traditional HPC modeling & simulation workloads, and AI data intensive workloads
- Customers leverage Dell expertise to reduce risk and expedite development, securing an optimal design with good ROI and low TCO

Usages in Finance segments worldwide spans many banking and insurance use cases. Benchmarked against STAC-A2 to measure performance on Monte Carlo simulations (SUT ID NVDA221007).

Workload-optimized rack-level system based on building block options provide flexibility in scalable design architecture

- PowerEdge Servers – XE8545
- GPU – 4x Nvidia A100 GPUs SXM4 40GiB
- Network Fabrics – Ethernet, Nvidia InfiniBand, Cornelis Networks OmniPath Express
- Dell Storage Solutions – Dell HPC BeeGFS, PowerScale F600 or F900, and Kalray Pixstor
- Deployment Software – Omnia, Bright Cluster Manager
- Support & Deployment Services

Dell Validated Design (DVD) for HPC Financial Services - Machine Learning

(Launched May 2023)

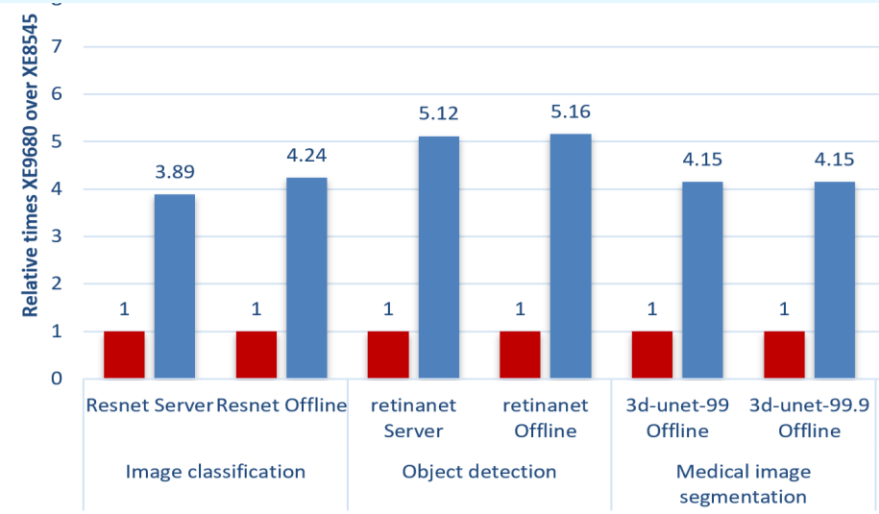
- Scalable Rack Units for compute and accelerated compute serve as baselines to architect scalable custom clusters
- Combine Scalable Rack Units to support compute-centric workloads, data-centric workloads, or both with the same cluster
- Validated by Dell with benchmarking for traditional HPC modeling & simulation workloads, and AI data intensive workloads
- Customers leverage Dell expertise to reduce risk and expedite development, securing an optimal design with good ROI and low TCO

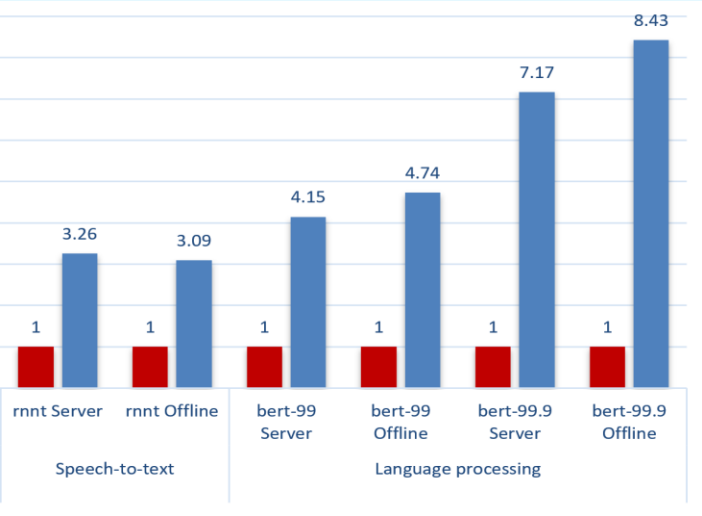
Benchmark against the STAC-ML benchmarking is forthcoming.

Workload-optimized rack-level system based on building block options provide flexibility in scalable design architecture

- PowerEdge Servers – R7625 or R760
- GPU – 1x NVIDIA H100-PCIe-80GB
- Network Fabrics – Ethernet, Nvidia InfiniBand, Cornelis Networks OmniPath Express
- Dell Storage Solutions – Dell HPC BeeGFS, PowerScale and Kalray pixstor
- Deployment Software – Omnia, Bright Cluster Manager
- Support & Deployment Services

General AI Performance Improvement 16G vs 15G





Dell HPC & AI Innovation Lab

3 clusters housed in a 13,000 ft^2 data center in Austin, TX

- Customer access to thousands of Dell servers, sophisticated storage and network systems for HPC
- Staffed with computer scientists, engineers and subject matter experts in a variety of disciplines available to help customers test and tune advanced computing solutions for their specific needs
- Help customers achieve faster time to results by shortening design cycle and configuration time

Many team members maintain security clearances to work on classified government projects

The Dell Technologies HPC & AI Innovation Lab team stays on the cutting edge, testing new technologies and tuning algorithms and applications to help customers keep pace with the constantly evolving landscape

- AI related computing paradigms are emerging quickly, and not many organizations have had the time to develop the skills required to design, deploy and manage them

Not STAC benchmarks

PURPOSE-BUILT

Accelerate AI Outcomes

PowerEdge XE9680

PowerEdge XE9640

PowerEdge XE8640

POWERED BY 4TH GENERATION INTEL XEON SCALABLE PROCESSORS

No-compromise Accelerated AI

- 8x NVIDIA H100 SXM5 700W 80GB NVLink GPUs or
- 8x NVIDIA A100 SXM4 500W 80GB NVLink GPUs
- Full NVLINK interconnectivity
- Air cooled operation (up to 35C)

Dense Acceleration

- 4x Intel Data Center Max Series GPU with GPU-GPU connectivity
- Dell Smart Cooled DLC GPUs
- 1:1 GPU-I/O enables faster data operations

Superior Performance

- 4x NVIDIA H100 SXM5 700W 80GB NVLink GPUs
- Full NVLINK interconnectivity
- GPU Direct Storage for fast data intake

New 16G Server Offering

More products at

https://www.delltechnologies.com/asset/en-us/products/servers/briefs-summaries/dell-poweredge-ai-servers-and-accelerators.pdf.external

Dell PowerEdge Acceleration-optimized servers for AI & HPC

DELL Technologies

Quick Reference Guide: Accelerators for PowerEdge servers

Businesses leaders today are faced with more data, decisions, and challenges than ever before, to grow the business and transform their operations for continued success. Faster decision-making, faster insights, faster real-time analysis are vital. Artificial Intelligence is the key, to helping businesses transform their data into insights and actionable results.

To design an infrastructure to deliver the capabilities which can make organizations successful with AI, the new potential of Generative AI and other demanding workloads requires a modern architecture approach where one of the biggest innovations is improved performance by accelerating insights, enabling a secure foundation for AI operations and the AI lifecycle with a trusted AI approach, and the addition of dense acceleration at scale for simplified operations and democratized AI.

Dell helps you put AI to work for you anywhere in any way to fast-track innovation, powering your AI workloads with accelerated insights, all from the new PowerEdge XE servers, powered by the Intel® Xeon® Scalable processors. Dell PowerEdge helps unleash your AI advantage with a modern compute foundation that continuously accelerates your journey, streamlining your operations securely.

	PowerEdge XE servers are Acceleration-Optimized, purpose-built for complex compute and AI/ML/DL and HPC intensive workloads.				PowerEdge rack servers are flexible, mainstream computing foundations for a wide range of applications, use cases and workloads.				
Rack Server	**XE9680**	**XE9640**	**XE8640**	**XE8545**	**R760xa**	**R750xa**	**R940xa**	**R750 / 7525 / 7515, R650 / 6525 / 6515**	**XR12**
									
Specifications	**No-compromise accelerated AI**	**Dense acceleration**	**Purpose-built performance**	**Superior outcomes for AI**	**Purpose-built flexibility**	**Purpose-built flexibility**	**Extreme acceleration**	**Mainstream performance**	**Edge performance**
Processor	Two 4th Generation Intel® Xeon® Scalable processors	Two 4th Generation Intel® Xeon® Scalable processors	Two 4th Generation Intel® Xeon® Scalable processors	Two 3rd generation AMD EPYC™ processors	Two 4th Generation Intel® Xeon® Scalable processors	Two 3rd Generation Intel® Xeon® Scalable processors	Up to four 2nd Generation Intel® Xeon® Scalable processors	One or Two 3rd Generation Intel® Xeon® Scalable or 3rd Generation AMD EPYC™ processors	One 3rd Generation Intel® Xeon® Scalable processor
Memory	32 DDR5 DIMMs, 4TB max	32 DDR5 DIMMs, 4TB max	32 DDR5 DIMMs, 4TB max	32 DDR4 DIMMs, 2TB max	32 DDR5 DIMMs, 4TB max	32 DDR4 DIMMs, 4TB max	48 DDR4 DIMMS, 12TB max	Up to 32 DDR4 DIMMs, 4TB max	8 DDR4 DIMMs, 1TB max
GPU support	8x NVIDIA(R) H100 Tensor Core SXM5 or 8x NVIDIA A100 SXM4 NVLink connectivity	4x Intel Data Center Max 1550 OAM GPU GPU-GPU connectivity	4x NVIDIA H100 Tensor Core SXM5 NVLink connectivity	4x NVIDIA A100 Tensor Core SXM4 NVLink connectivity	4 x 350W Double-Wide or 12 x 75W Single-Wide	4 x 300W Double-Wide or 6 x 75W Single-Wide	4 x 300W Double-Wide	Up to 3 x 300W Double-Wide or 6 x 75W Single-Wide	Up to 2 x 300W Double-Wide or Single-Wide
Other features	Air-cooled operation (up to 35C) 6U rack height Up to 8 x 2.5" drives Up to 10 x PCIe Gen5	Liquid-cooled CPU and GPU operation 2U rack height Up to 4 x 2.5" drives 2 x PCIe Gen5	Air-cooled operation (up to 35C) 4U rack height Up to 8 x 2.5" drives Up to 4 x PCIe Gen5	Air-cooled operation (up to 35C) 4U rack height Up to 10 x 2.5" drives Up to 4 x PCIe Gen4	Air-cooled operation (up to 35C) 2U rack height Up to 8 x 2.5" drives Up to 4 x PCIe Gen5	Air-cooled operation (up to 35C) 2U rack height Up to 8 x 2.5" drives Up to 4 x PCIe Gen4	Air-cooled operation (up to 35C) 4U rack height Up to 24 x 2.5" drives Up to 12 x PCIe Gen3	Air-cooled operation (up to 35C) 1U or 2U rack height" Up to 8 x 2.5" drives Up to 8 x PCIe Gen4	*-5°C to 55°C 2U rack height Up to 4 x PCIe Gen4
Applications and use-cases	Large data set language models, Natural Language Processing, AI ML DL Training, HPC, CRISP, Healthcare, CSP/HPCaaS, Finance, Academia, Generative AI/GPT	AI ML DL Training, HPC, Modeling & Simulation, Healthcare, Life Sciences, Finance	Medium data set language Models, Modeling & Simulation, AI, ML/DL Training and Inferencing	AI ML Training and inferencing, small and medium data set langauge models	AI & ML training and inferencing, data analytics, HPC, VDI & Performance graphics	AI & ML training and inferencing, data analytics, HPC, VDI & Performance graphics	GPU database acceleration, data analytics, AI, machine learning	Light duty AI/ML/DL training, inferencing, VDI, Performance graphics, Edge	Edge AI training, Inferencing, Telco, rendering/modeling
Availability	now	1H, 2023	1H, 2023	now	1H, 2023	now	now	now	now

DELL Technologies

Sampling of more AI at Dell, not just for HPC!

- https://www.dell.com/en-us/dt/solutions/artificial-intelligence/index.htm#accordion0&tab0=0
- https://www.dell.com/en-us/blog/categories/artificial-intelligence/
- https://www.dell.com/en-us/dt/services/solutions/artificial-intelligence-services.htm
- https://www.dell.com/en-us/dt/ai-technologies/index.htm
- https://dell.sharepoint.com/sites/apj-isg-coc/Shared%20Documents/APJ%20Server%20GPU%20Webinar%20Series/FY23%20APJ%20Server%20GPU%20webinar%20series/FY23Q2%20APJ%20Server%20GPU%20webinar%20series/NVIDIA%20LaunchPad%20and%20NVAIE-June%202022.pdf
- https://www.dell.com/en-us/dt/corporate/newsroom/announcements/detailpage.press-releases~usa~2023~05~dell-technologies-and-nvidia-introduce-project-helix-for-secure%2c-on-premises-generative-ai.9726.htm#/filter-on/Country:en-us
- Go to Dell Technologies and search for more!