Optimize and Accelerate AI Inferencing in Real-World Environments

In collaboration with: intel

Tech Note by

Todd Mottershead
Todd.mottershead@dell.com
Seamus Jones
Seamus.jones@dell.com
Lokendra Uppuluri
Lokendra.Uppuluri@intel.com
Krzysztof Cieplucha
krzysztof.cieplucha@intel.com

Summary

There are multiple considerations to take into account when deploying artificial intelligence environments. This paper serves as a discussion and suggestion as to the possible hardware configurations to achieve a server infrastructure deployment that is secure and can grow with your increased need based on the most recent 15th Generation PowerEdge Server portfolio offerings.

Reliable and fast access to data is increasingly critical for every

Businesses are increasingly using artificial intelligence (AI) to increase revenue, drive operational efficiencies and innovate for new products. AI use cases powered by deep learning (DL) generate some of the most powerful insights. But meeting the required service-level agreements (SLAs) in a production environment can be challenging, especially in a microservices architecture.

Dell EMC™ PowerEdge™ servers are built on an Intel® architecture that has been benchmark tested and verified for demanding DL workloads. This reference architecture features 3rd Generation Intel® Xeon® Scalable processors, Intel® Optane™ Solid State Drives (SSDs), and Intel® TLC 3D NAND SSDs for high performance in real-world scenarios. While discrete, hardware-based AI accelerators can be used in extreme DL use cases, 3rd Generation Intel Xeon Scalable processors remain the only data center CPUs with built-in AI acceleration, support for end-to-end data-science tools and an ecosystem of smart solutions. 3rd Generation Intel Xeon Scalable processors deliver 1.5x more performance than other CPUs across 20 popular machine learning (ML) and DL workloads.[i]

Two configurations are available: the Base configuration provides a balance between price, performance and built-in Intel technologies that enhance performance and efficiency for inferencing on AI models. The Plus configuration is designed for even faster AI inferencing.

Key Considerations

- **Accelerated inferencing.** Intel performance optimizations that are built into Dell EMC PowerEdge servers can help speed AI inferencing. These optimizations include Intel and open-source software and hardware technologies, such as Intel® Deep Learning Boost (Intel® DL Boost) with Vector Neural Network Instructions (VNNI) for AI acceleration, Intel® oneAPI Deep Neural Network Library (Intel® oneDNN), the OpenVINO™ toolkit and optimized versions of TensorFlow™ and PyTorch®.

- **Scalability.** Dell EMC PowerEdge configurations are built to scale so that IT can deploy AI inferencing in production environments quickly and efficiently.

- **Simplified stack.** Dell EMC PowerEdge servers built on Intel architecture are a turnkey solution in an optimized, pre-tuned and tested configuration for low-latency, high-throughput inferencing.

Available Configurations

	Base Configuration	Plus Configuration
Platform	Dell EMC™ PowerEdge™ R650, supporting up to 10 NVM Express® (NVMe®) drives (direct connection with no Dell™ PowerEdge RAID Controller [PERC]), 1 RU	
CPU	2 x Intel® Xeon® Gold 6348 processor (28 cores at 2.6 GHz)	2 x Intel® Xeon® 8368 processor (38 cores at 2.4 GHz)
DRAM	256 GB (16 x 16 GB DDR4-3200)	512 GB (16 x 32 GB DDR4-3200)
Boot device	Dell EMC™ Boot Optimized Server Storage (BOSS)-S2 with 2 x 480 GB Intel® SSD S4510 M.2 Serial ATA (SATA) (RAID1)	
Storage adapter (optional)	Dell PERC H755N front NVMe RAID adapter[ii]	
Cache storage (optional)	1 x 400 GB Intel® Optane™ SSD P5800X (PCIe Gen4) or 1 x 375 GB Intel Optane SSD DC P4800X (PCIe Gen3)[iii]	
Capacity storage	1 x (up to 9 x) 3.84 TB Intel SSD DC P5500 (PCIe Gen4, read-intensive)	
Network interface controller (NIC)	Intel® Ethernet Network Adapter E810-XXV for OCP3 (dual-port 25 Gb)	

Learn More

Contact your Dell or Intel account team for a customized quote 1-877-289-3355

Read the solution brief “Intel Select Solutions for AI Inferencing”
https://www.intel.com/content/www/us/en/artificial-intelligence/solutions/select-solution-for-ai-inferencing.html

Visit the Dell Technologies HPC & AI Innovation Lab: www.delltechnologies.com/en-us/solutions/high-performance-computing/HPC-AI-Innovation-Lab.htm

[i] Source: Claim 43 at: Intel. “3rd Generation Intel Xeon Scalable Processors – Performance Index.” www.intel.com/3gen-xeon-config.
[ii] An NVMe RAID adapter is optional but recommended for configurations with a large number of capacity drives.
[iii] The Intel® Optane™ SSD P5800X is recommended, but the previous-generation Intel Optane SSD DC P4800X can be used instead if the Intel Optane SSD P5800X is not yet available.