Dell AI Factory with NVIDIA Solution ID 19845070.1

About this solution:

This configuration contains 8 PowerEdge XE9680 servers, 5 PowerScale F710 nodes, and 64 NVIDIA H200SXM GPUs. These systems are capable of powering a variety of generative AI workloads, including large language model (LLM) inferencing and training, as well as 3D graphics, rendering, and video.

The Dell AI Factory with NVIDIA is designed to simplify development and accelerate AI adoption with a full-stack offering—encompassing computing, networking, storage, software, and services.

Learn more about Dell AI Factory with NVIDIA at https://www.dell.com/en-us/lp/nvidia-ai

List of components:

Rack-Scale Hardware:

Please Note: Actual rack configurations will vary based on power per rack. This solution is presented in a rack that can accommodate approximately 48/2kWs.

- 3x Dell PowerEdge R660
- 8x Dell PowerEdge XE9680
  - Dual Intel Platinum CPU/32c/ 2.112TB H200SXM RAM
  - 64x H200SXM GPU (Total 5.1TB)
  - NVIDIA Bluefield-3 (1x3220/8x3140H)
- 3x NVIDIA Spectrum™-4 SN5600 400G BE + 200G FE
- 2x PowerSwitch S5232F-ON Storage Cluster 100G BE
- 1x NVIDIA SN2201 ToR
- 5x PowerScale F710 (Total Storage 284 TB)
- 3x APC 750x1200 42U Rack (A6921473)
- 10x PDU (AC021024)

Software:

- 11x Ubuntu OS
- Upstream Kubernetes
- 67x NVAIE
- (NVIDIA NeMo Microservices for Generative AI, NIMs, RAG Systems, BCMe)

Dell Technologies Services:

- ProSupport+ or ProSupport One
- ProDeploy
- ProConsult