SOLUTION BRIEF

DELIVERING NVIDIA AI ENTERPRISE ON RED HAT® OPENSHIFT®

Supermicro NVIDIA-Certified Systems, with 4th Gen Intel® Xeon® Scalable Processors

SYS-221HE-FTNR

SYS-221H-TNR

SYS-421GE-TNRT

SYS-821GE-TNHR

TABLE OF CONTENTS

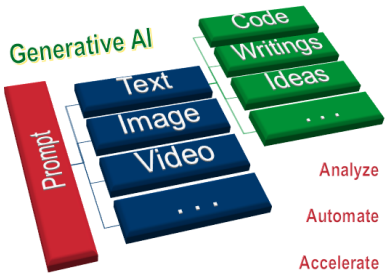
SUPERMICRO

As a global leader in high performance, high efficiency server technology and innovation, we develop and provide end-to-end green computing solutions to the data center, cloud computing, enterprise IT, big data, HPC, and embedded markets. Our Building Block Solutions® approach allows us to provide a broad range of SKUs, and enables us to build and deliver application-optimized solutions based upon your requirements.

Executive Summary

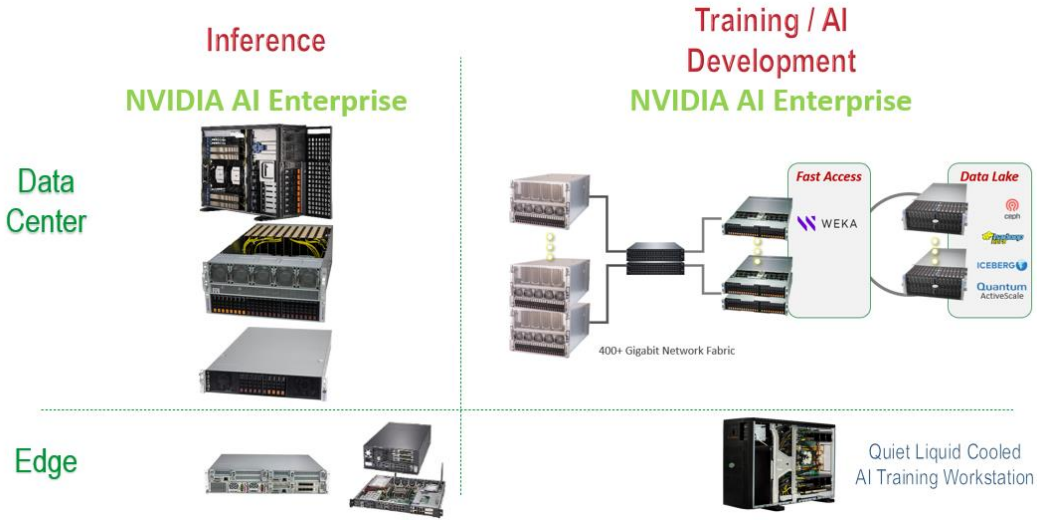
AI is a game changer for many businesses. With multitudes of mature, pre-trained AI models, including Generative AI models, businesses can deploy AI to analyze data quickly to identify issues and opportunities, to automate interactions with customers, partners, and suppliers, and to accelerate content creation and product development.

Supermicro offers a complete line of time-to-market systems supporting a wide range of NVIDIA GPUs. These run the NVIDIA AI Enterprise software suite, which enables rapid AI development and deployment. Red Hat® OpenShift® provides a reliable environment to support MLops workflows. Supermicro accelerates AI implementations by delivering systems with OpenShift running on the latest generations of Intel CPUs and NVIDIA AI Enterprise running on NVIDIA GPUs. As a result, customers can quickly take advantage of the game changing AI capabilities.



AI Training

AI training for small AI models, such as image and object recognition, can be accomplished using one or several GPU systems. Large language models (LLM) require one or multiple racks of these systems. The number of servers and the amount of time can be reduced using pre-trained models provided by NVIDIA AI Enterprise. Supermicro offers very fast NVMe based storage systems to enable fast GPU-Direct access to data to train the AI models.

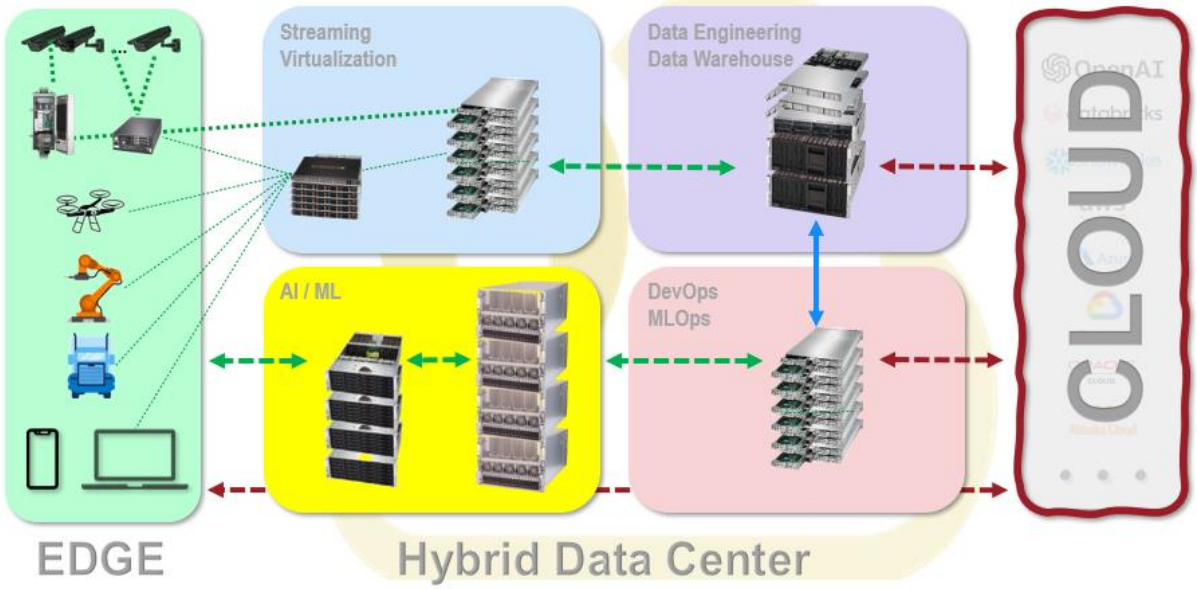


AI Inference

After the AI models are trained, AI inference can be done in the data center or on the edge. The inference servers can automatically deploy Trained AI models using TensorRT and NVIDIA Inference Server, available from NVIDIA AI Enterprise.

AI Data Flow and Workflow

Comprehensive data flow and workflow can be incorporated as part of the business. Supermicro offers systems to collect data at the edge. Data are streamed to the hybrid data center and consolidated into data lakes. Data are then cleansed and formatted for AI training. After training, the trained AI models can be automatically exported to inference servers. Supermicro offers systems and storage to support different aspects of AI data flow and workflow.



SUPERMICRO NVIDIA CERTIFIED SERVERS

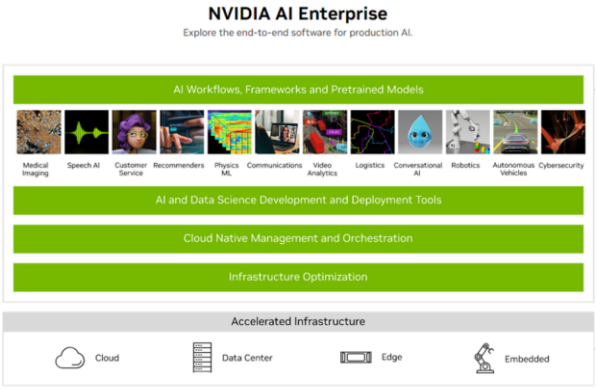
Supermicro servers with NVIDIA GPUs are NVIDIA certified. H100, A100, L40, L4, and other GPUs are supported. Choices of Intel® Xeon® Scalable processors, system memory up to 8TB, PCIe Gen4 or Gen 5 connectivity, NVMe drives, 400Gbit/s network connectivity, redundant power IPMI/Redfish management, TPM 2.0, hardware Root of Trust security. NVIDIA certifications for the 4th Xeon Scalable systems with H100 are in progress.

Red Hat OpenShift

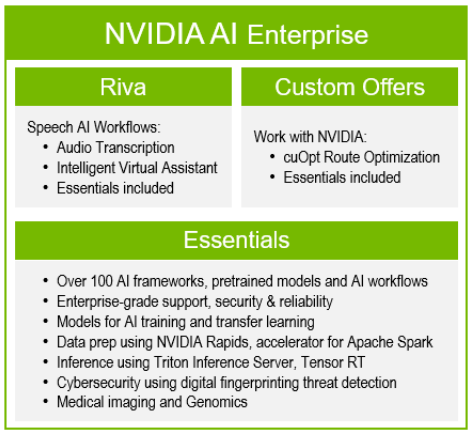
Red Hat® OpenShift® is an enterprise-ready Kubernetes container platform built for an open hybrid cloud strategy. It provides a consistent application platform to manage hybrid cloud, multi-cloud, and edge deployments. Using GPU Operators and other Operators, Red Hat OpenShift enables easy setup and robust operations running NVIDIA AI Enterprise workloads.

NVIDIA AI Enterprise Software Suite

The NVIDIA AI Enterprise software suite includes AI tools and frameworks, cloud native deployment, and infrastructure optimization software to enable rapid AI development and deployment. The software suite is offered with the Essentials version to support 100 AI frameworks and many AI deployments, along with premium versions: Riva to support speech AI and custom to support specialized AI deployment.



By providing minimal risk and a simple approach to integrating AI into the existing enterprise container environment, NVIDIA AI Enterprise enables an end-to-end software stack approach to start using AI in the enterprise. Enterprise developers can initially run small trials until they feel comfortable expanding to more extensive deployment. At that point, the solution is very scalable to deployment in multiple racks.



Enterprise Support

Services

Using the NVIDIA AI Enterprise software suite, enterprise customers get enterprise-grade support for the entire system, from AI software to the virtualization and system hardware, including NVIDIA data center GPUs and network accelerators optimized in Supermicro systems. As a solutions provider, Supermicro offers and supports the entire Supermicro systems with Red Hat OpenShift and NVIDIA AI Enterprise software.

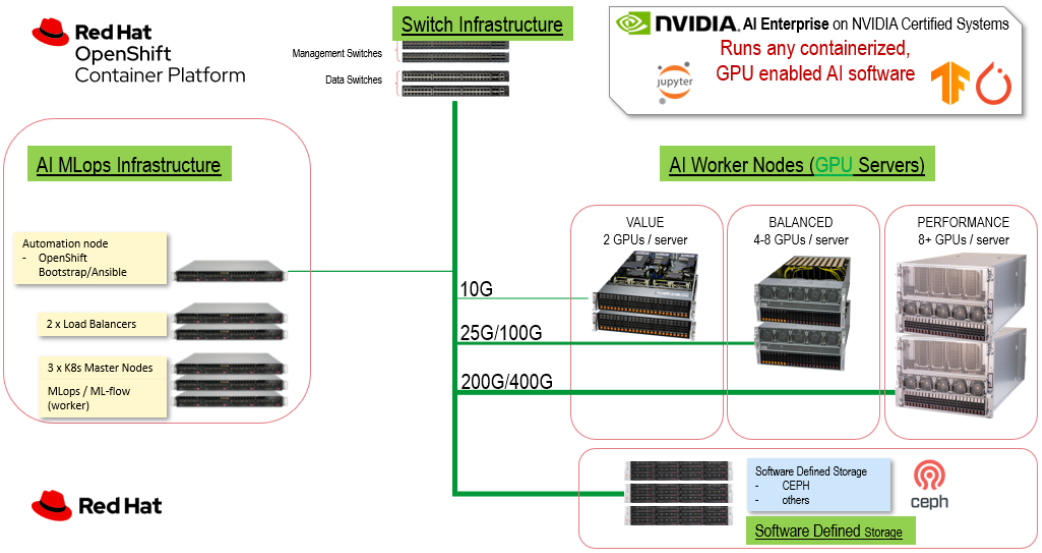
Management & Security

Supermicro systems provide out-of-band and in-band monitoring. Using out-of-band IPMI and Redfish management, the health and operation of each server in the cluster can be managed. The servers also come with optional TPM 2.0 and Root of Trust security features.

Supermicro Reference Architecture for NVIDIA AI Enterprise running Red Hat OpenShift

Supermicro Reference Architecture for NVIDIA AI Enterprise and Red Hat OpenShift provides a scalable architecture. As a result, enterprise AI developers can quickly develop AI solutions to increase efficiency and enable new services using pre-trained AI models. Supermicro accelerates the deployment of AI containers in the Red Hat OpenShift's orchestrated container environment with the help of Generative AI to automate tested installation scripts. Enterprise support is available on the Supermicro systems that are NVIDIA-Certified, and Red Hat certified, including the entire software stack.

	Red Hat OpenShift Master Nodes, SuperCloud Composer Node	Edge AI Worker	Small AI Worker	Medium AI Worker	Large AI Worker
Server	SYS-110P-WTR 1U	SYS-221HE-FTNR 2U	SYS-221H-TNR 2U	SYS-421GE-TNRT 4U	SYS-821GE-TNHR 8U
Number of Servers	4	1 to 256	1 to 256	1 to 256	1 to 256 (per POD)
Server Configuration	1 x Xeon Scalable 4310 (12 core) 64GB 256GB M.2 2 x 1TB SSD Dual 10GbE	2 x Xeon Scalable 5418Y (24 core) 256GB 256GB M.2 2 x 1TB SSD Dual 10GbE	2 x Xeon Scalable 6442Y (24 core) 256GB 256GB M.2 2 x 1TB SSD Dual 25GbE	2 x Xeon Scalable 6430 (32 core) 1024GB 2 x 1TB M.2 2 x 4TB SSD 4 x 200GbE	2 x Xeon Scalable 8468 (48 core) 2048GB 2 x 1TB M.2 2 x 1TB SSD 8 x 400GbE
GPU	-	1 to 3 x A30, A100, H100	1 to 4 x A30, A100, H100	1 to 8 x A100 or H100	HGX-H100 8-GPU
BMC Switches (per 32 worker nodes)	-	-	2 x SSE-G3648B	2 x SSE-G3648B	2 x SSE-G3648B
Data Switches (per 32 worker nodes)	-	-	2 x SSE-X3648SR	2 x SSE-SN3420-CB2RC or 2 x SSE-SN3700-CS2RC	2 x SSE-SN3700-CS2RC or 2 x SSE-SN3700-VS2RC



Example Applications

Here are example applications using containerized machine learning infrastructure. Specific customer solutions need to be adjusted to match customer needs.

	Number of Simultaneous Users	CPU Cores	System Memory	Storage	NVIDIA GPU	GPU System
AI Development – Smaller Jobs	Up to 12	48	256GB	100TB	2 x A30 per node	SYS-221H-TNR
AI Development – Medium Jobs	Up to 100	64	1024GB	200TB	4 x H100	4 x SYS-421GE-TNRT
AI Development – Large Jobs (multiple GPUs)	Multiple 100's	96	2048GB	400TB	8 x H100	8 x SYS-821GE-TNHR
AI Inference	Continual	64	1024GB	100TB	4 x A100	2 x SYS-421GE-TNRT

Conclusion

Supermicro NVIDIA-Certified Systems support NVIDIA AI Enterprise running on Red Hat OpenShift to enable AI developments and delivery to run small to large AI workloads. The reference architectures with specific small, medium, and large configurations provide a robust framework for customers to start using NVIDIA AI Enterprise, running in a robust orchestrated container environment provided by Red Hat OpenShift.

Supermicro offers these as integrated solutions, including systems, software, and support. Please call your Supermicro representative for more information.

References

1. NVIDIA AI Enterprise
2. Red Hat OpenShift
3. Supermicro NVIDIA-Certified Systems
4. Supermicro GPU Servers using Intel Xeon Scalable processors